I See What You Did There: Towards a Gaze Mechanism for Joint Actions in Human-Robot Interaction

Michael Koller, Astrid Weiss, Markus Vincze

Abstract

We imagine that service robots must collaborate with humans in physical object manipulation tasks to be of assistance in everyday scenarios, such as setting a table. This collaboration requires the capability of joint attention to smoothly accomplish a shared goal. One special modality for joint attention is the gaze behavior of an actor. Herein, we discuss the human gaze in physical tasks and its underlying cognitive mechanisms, a novel probabilistic robotic gaze controller in object-centred collaborative physical tasks, and its inclusion in a well-known joint action human-robot interaction (HRI) benchmark. First, 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 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 gaze direction can be interpreted by the interaction partner. 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 results of a pilot study.

Keywords

human-robot interaction, joint attention, joint action, gaze, eye-tracking

1 Introduction

Think of a situation where you have 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 to squeeze the sofa around the tight corner up 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 just 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

https://doi.org/10.34727/2022/isbn.978-3-85448-052-5_7

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 chapter, we first contribute a discussion of results in psychology related to this topic. Specifically, we review research on joint attention [Baron-Cohen 1994; Mundy and Newell 2007] and theory of mind [Baron-Cohen 1997] 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. 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 Clodic et al. [2017], based on building a tower out of various wooden blocks. 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 and discuss how to include it into the joint action benchmark. Our contribution extends a stochastic gaze controller for static scenarios to dynamic ones.

2 Joint Attention in Psychology

Joint attention has been studied since the 1970ies [Scaife and Bruner 1975]. 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 [Krämer et al. 2011] 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 [Akhtar and Gernsbacher 2007]. One of the earliest reports of joint attention appeared 1975 in an article by Scaife and Bruner [1975] 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.

2.1 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 Theory of Mind (ToM). Tomasello [1995] 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)." [Tomasello 1995, p. 105]

The term "theory of mind" was coined by Premack and Woodruff [1978] 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)."[Tomasello 1995, p. 104] [1]

Baron-Cohen [1994, 1997] 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."

1 Although the term *joint attention* originated in developmental psychology, other approaches in psychology also provided results on the topic, some of which is 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.

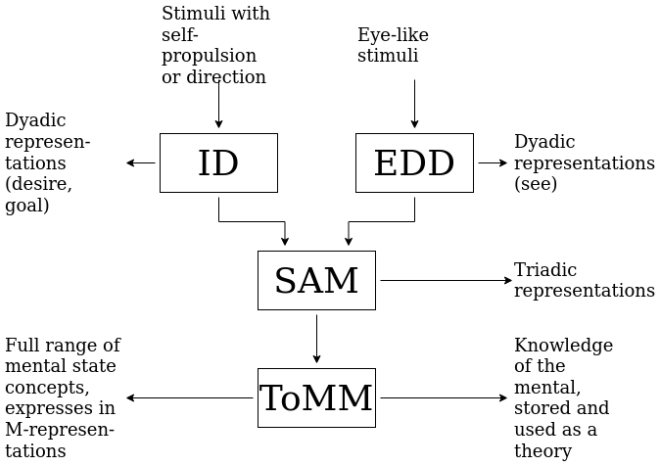

Figure 1 Mind-reading system, adapted from Baron-Cohen [1994].

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) (Figure 1). 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 [Langton et al. 2000]. 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).

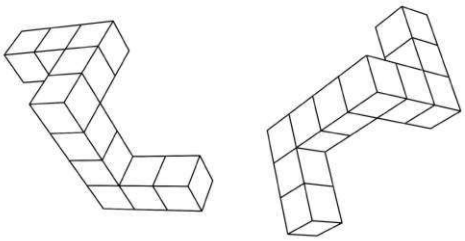

Figure 2 An example of a mental rotation task, adapted from [Just and Carpenter 1976].

2.2 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 *responding to joint attention* (RJA) and *initiating joint attention* (IJA) [Mundy and Newell 2007]. 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 [Tomasello et al. 2005].

2.3 Eye-Mind Hypothesis

The gaze occurs first to gather information, while it also signals information to observers, either intentionally or unintentionally. Just and Carpenter [1976] 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 (Figure 2) 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" [Just and Carpenter 1976,

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: Webb and Renshaw [2008] 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.

2.4 Types of Gaze Behavior

As discussed in previous sections, there is strong evidence of some connection of 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. Kaplan and Hafner [2006] 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 [Andrist et al. 2014].

3 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 [Kolbeinsson et al. 2019]; in other words, joint attention as employed in the sofa moving example mentioned in the introduction. In our work, we explicitly focus on human-robot interaction (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, Krämer et al. [2011] 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 artificial intelligence (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.

3.1 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. Imai et al. [2003] proposed an HRI joint attention mechanism in 2003. 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 looked more frequently at the poster, when the robot displayed the proposed attention mechanism.

Huang and Thomaz [2010, 2011] 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 while giving directions. The overall result indicates that robots with their joint attention implementation yielded better results in the responding to pointing actions task, and were considered more nat-

ural in the video-based experiments. Huang and Thomaz [2010] mentioned, that it is unclear how to design the specific timings of the EJA component.

Pereira et al. [2019] 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 [Frankel et al. 2012], but also HRI. Grigore et al. [2013] 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, Moon et al. [2014] 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 occurs sometimes, 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 Planning for Joint Human-Robot Interaction

As Baron-Cohen [1994] 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 opening 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 as 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* [Harnad 1990; Coradeschi et al. 2013]. 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 [Mirsky et al. 2021]) 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 ab-

stract 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 [LaValle 2006]. 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 [LaValle 2006]. Thus, each action outcome is assigned an 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 [LaValle 2006], where agents repeatedly chose actions according to their current state. A single agent MDP is defined by 1) a non-empty ***state space X***, which is a finite or countably infinite set of states; 2) for each $x \in X$ a ***finite, non-empty action space U(x)*** with a ***termination action*** (it is applied when reaching a goal state); 3) a finite, non-empty ***nature action space*** $\Theta(x, u)$ for each $x \in X$ and $u \in U(x)$ (a ***nature*** decision maker represents uncertainty in the action outcome); 4) a state transition function f that produces a state, $f(x, u, \theta)$, for every $x \in X$, $u \in U$, and $\theta \in \Theta(x, u)$; 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 $x_I \in X$; 7) a goal set $X_G \subset X$, 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 $\pi : X \rightarrow U$ must be found for all $x \in X$ 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 1). $X_1, X_2, \ldots, X_t$ denotes the sequence of random variables up to timestep t, where the outcomes are $x_i \in X$. 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.

$$Pr(X_{t+1}) = x_{t+1} | X_1 = x_1, X_2 = x_2, \ldots, X_t = x_t) = Pr(X_{t+1}) = x_{t+1} | X_t = x_t) \tag{1}$$

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 anoher 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 $x_t \in X$, but holds a ***belief*** about the current state, expressed as a probability over X. Including this belief into planning lifts the problem formulations from the state space into the state belief space.[2]

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* [LaValle 2006]. 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 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 function. 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 wide-spread 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

2 Literature presented in this chapter as well as our contribution only concerns planning in state space.

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 [Ghallab et al. 2016].

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 planning domain `blocksworld` in the *Planning Domain Definition Language* [Fox and Long 2003] (PDDL), 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) [Lifschitz 1987]. 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.

3.3 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 [Ramírez and Geffner 2009, 2010; Sohrabi et al. 2016]. 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. 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.

3.4 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 [Johnson et al. 2009; Schulz et al. 2018; Barchard et al. 2020; Jensen 2021].

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. Devin et al. [2017] 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[3] (Figure 5, 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. Howeer, 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 (Figure 3). 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.

3 https://robots.ieee.org/robots/pr2/, Image source: https://www.wevolver.com/wevolver.staff/pr2

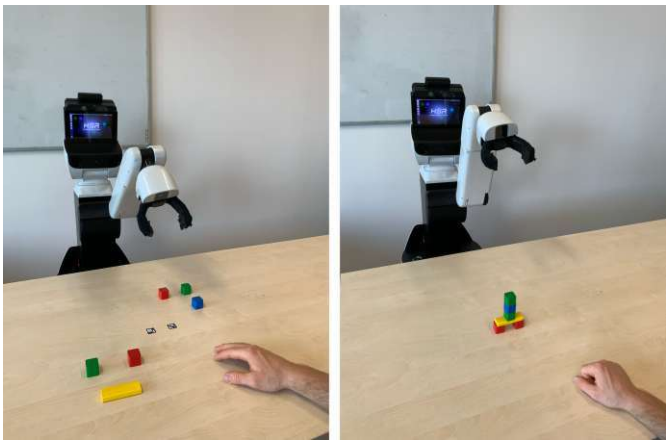

Figure 3 Joint action task described in [Devin et al. 2017]. Left: Initial configuration. Right: Goal State.

The authors approach this scenario as a multi-agent planning problem. The robot 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 actors 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, Clodic et al. [2017] propose a joint action scenario similar to Devin et al. [2017]. 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 [...]?” [Clodic et al. 2017, p. 2] The proposed simple HRI scenario has the following setup and assumptions:

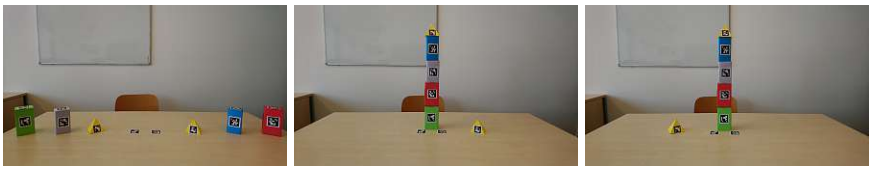

Figure 4 Left: Initial configuration. Middle and Right: The two possible goal states.

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

Figure 4 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.

4 Toward a Gaze Mechanism for Joint Actions

As described above, one of the two core questions posed by Clodic et al. [2017] 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[4] (HSR) (Figure 5) 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[5] or Nao[6] (Figure 6) have the same degrees of freedom in their heads and have al-

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

5 https://www.softbankrobotics.com/emea/en/pepper

Figure 5 Two domestic service robots. Left: Toyota Human Support Robot (HSR). Right: PR2 by WillowGarage.

ready been used in gaze related HRI studies. Research has shown that their head orientation communicates attention [Breazeal et al. 2005; Takayama et al. 2011] 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 Admoni and Scassellati [2017].

4.1 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 [Hayhoe and Ballard 2005; Pelz et al. 2001]. 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, are 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 se-

6 https://www.softbankrobotics.com/emea/en/nao

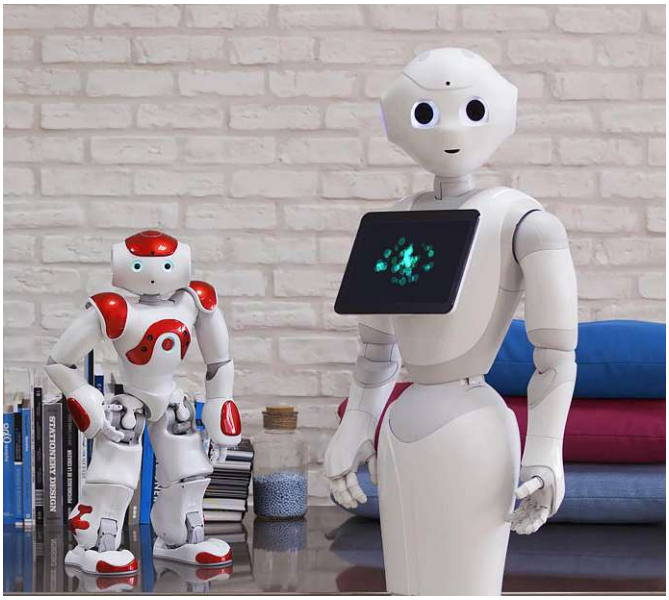

Figure 6 Two social humanoid robots by Softbank Robotics. Left: Nao. Right: Pepper.

quence 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.

4.2 Modeling the Sequence of Gaze Targets

Next, we discuss how to create a gaze model for the above-mentioned tasks. Lehmann et al. [2017]; Acarturk et al. [2021] 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[7] 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 directions).

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 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 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 Andrist et al. [2014], 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. Provid-

7 https://www.tobiipro.com/product-listing/tobii-pro-glasses-2/

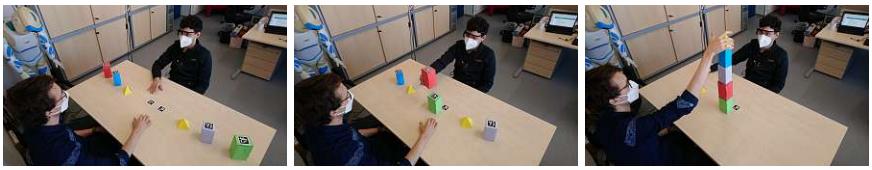

Figure 7 Gaze data capturing during the pilot study. Left: Initial position. Middle: Eye-tracked participant places a block from the reachable area. Right: Placement of the pyramid block. Both participants can place their pyramid, and after a negotiation phase, the other participant places the final piece.

ing a gaze controller for the joint action benchmark task described earlier is thus helpful to handle more realistic scenarios in the future.

5 Data Collection for our Stochastic Gaze Control

We describe how to adapt the procedures from Lehmann et al. [2017]; Acarturk et al. [2021] to a collaborative object manipulation task. In a pilot study, we recreated the block stacking task with the pyramid top presented in Clodic et al. [2017] (Figure 7). Two human participants sit opposite each other at a table. One of the two participants per trial wore a PupilLabs Core[8] [Kassner et al. 2014] 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 4 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 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.

8 https://pupil-labs.com/products/core/

The participants were informed that this is 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 in the future. 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 problms 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 to 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 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 occured either immediately or after a short period of inactivity.

For each gaze data sample, we conducted the following evaluation: Using fiducial markers[9], as well as (the partner’s) hand and face tracking [Lugaresi et al. 2019] 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 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 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 embod-

9 https://april.eecs.umich.edu/software/apriltag

	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 1 DTMC transition probabilities of eye-tracked locations.

iment.) and creating a gaze sequence by starting in a random or predetermined (e.g., face) state. The next state is always 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 whether the gaze behavior in the action phase (placement up to the last block) differs from in the negotiation phase (placement of either pyramid).

5.1 Creating a Gaze Controller for Time-Variant Scenarios

Table 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*

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

blocks group blocks that have been placed before *previous*, and must be placed after *next*, respectively. The controller in Table 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.

5.2 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 Clodic et al. [2017]. 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 Andrist et al. [2014].

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 Lehmann et al. [2017]; Acarturk et al. [2021], and Andrist et al. [2014] 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* [Aloimonos et al. 1988]. 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.

6 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.

Clearly, 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.

Bibliography

Cengiz Acarturk, Bipin Indurkya, Piotr Nawrocki, Bartlomiej Sniezynski, Mateusz Jarosz, and Kerem Alp Usal. 2021. Gaze aversion in conversational settings: An investigation based on mock job interview. *Journal of Eye Movement Research* 14, 1.

Henny Admoni and Brian Scassellati. 2017. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction* 6, 1, 25–63.

Nameera Akhtar and Morton Ann Gernsbacher. 2007. Joint attention and vocabulary development: A critical look. *Linguistics and Language Compass* 1, 3, 195–207.

John Aloimonos, Isaac Weiss, and Amit Bandyopadhyay. 1988. Active vision. *International Journal of Computer Vision* 1, 4, 333–356. https://doi.org/10.1007/BF00133571

Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. 2014. Conversational gaze aversion for humanlike robots. In *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 25–32.

Kimberly A Barchard, Leiszle Lapping-Carr, R Shane Westfall, Andrea Fink-Armold, Santosh Balajee Banisetty, and David Feil-Seifer. 2020. Measuring the perceived social intelligence of robots. *ACM Transactions on Human-Robot Interaction (THRI)* 9, 4, 1–29.

Simon Baron-Cohen. 1994. How to build a baby that can read minds: Cognitive mechanisms in mindreading. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition,* 13(5), 513–552.

Simon Baron-Cohen. 1997. *Mindblindness: An essay on autism and theory of mind.* Cambridge: MIT Press.

Cynthia Breazeal, Cory D Kidd, Andrea Lockerd Thomaz, Guy Hoffman, and Matt Berlin. 2005. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *2005 IEEE/RSJ international conference on intelligent robots and systems.* IEEE, 708–713.

Aurélie Clodic, Elisabeth Pacherie, Rachid Alami, and Raja Chatila. 2017. Key elements for human-robot joint action. In *Sociality and Normativity for Robots. Studies in the Philosophy of Sociality*, Hakli R, Seibt J (eds). Springer, Cham. 159–177 https://doi.org/10.1007/978-3-319-53133-5_8

Silvia Coradeschi, Amy Loutfi, and Britta Wrede. 2013. A short review of symbol grounding in robotic and intelligent systems. *KI - Künstliche Intelligenz* 27, 2, 129–136. https://doi.org/10.1007/s13218-013-0247-2

Sandra Devin, Aurélie Clodic, and Rachid Alami. 2017. About decisions during human-robot shared plan achievement: Who should act and how? *International Conference on Social Robotics*, 453– 463.

Maria Fox and Derek Long. 2003. PDDL2. 1: An extension to PDDL for expressing temporal planning domains. *Journal of Artificial Intelligence Research* 20, 61–124.

Richard M Frankel, Mindy Flanagan, Patricia Ebright, Alicia Bergman, Colleen M O'Brien, Zamal Franks, Andrew Allen, Angela Harris, and Jason J Saleem. 2012. Context, culture and (non-verbal) communication affect handover quality. *BMJ Quality & Safety* 21, Suppl 1, 121–128. DOI: 10.1136/bmjqs-2012-001482

Malik Ghallab, Dana Nau, and Paolo Traverso. 2016. *Automated planning and acting.* Cambridge University Press. https://doi.org/10.1017/CBO9781139583923

Elena Corina Grigore, Kerstin Eder, Anthony G Pipe, Chris Melhuish, and Ute Leonards. 2013. Joint action understanding improves robot-to-human object handover. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems.* IEEE, 4622–4629.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena* 42, 1-3, 335–346.

Mary Hayhoe and Dana Ballard. 2005. Eye movements in natural behavior. *Trends in Cognitive Sciences* 9, 4, 188–194.

Chien-Ming Huang and Andrea L Thomaz. 2010. Joint attention in human-robot interaction. In *2010 AAAI Fall Symposium Series.*

Chien-Ming Huang and Andrea L Thomaz. 2011. Effects of responding to, initiating and ensuring joint attention in human-robot interaction. *International Conference on Robot & Human Interactive Communication.* IEEE. 65–71.

Michita Imai, Tetsuo Ono, and Hiroshi Ishiguro. 2003. Physical relation and expression: Joint attention for human-robot interaction. *IEEE Transactions on Industrial Electronics* 50, 4, 636–643.

Alexander Birch Jensen. 2021. Towards Verifying a Blocks World for Teams GOAL Agent. *International Conference on Agents and Artificial Intelligence,* (1). 337–344.

Matthew Johnson, Catholijn Jonker, Birna van Riemsdijk, Paul J Feltovich, and Jeffrey M Bradshaw. 2009. Joint activity testbed: Blocks world for teams (BW4T). *International Workshop on Engineering Societies in the Agents World.* Springer. 254–256.

Marcel Adam Just and Patricia A Carpenter. 1976. Eye fixations and cognitive processes. *Journal of Cognitive Psychology* 8, 4, 441–480.

Frederic Kaplan and Verena V Hafner. 2006. The challenges of joint attention. *Journal of Interaction Studies* 7, 2, 135–169.

Moritz Kassner, William Patera, and Andreas Bulling. 2014. Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing.* ACM. 1151–1160. https://doi.org/10.1145/2638728.2641695

Ari Kolbeinsson, Erik Lagerstedt, and Jessica Lindblom. 2019. Foundation for a classification of collaboration levels for human-robot cooperation in manufacturing. *Production & Manufacturing Research* 7, 1, 448–471.

Nicole C Krämer, Sabrina Eimler, Astrid Von Der Pütten, and Sabine Payr. 2011. Theory of companions: what can theoretical models contribute to applications and understanding of human-robot interaction? *Journal of Applied Artificial Intelligence* 25, 6, 474–502.

Stephen R H Langton, Roger J Watt, and Vicki Bruce. 2000. Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Sciences* 4, 2, 50–59.

Steven M LaValle. 2006. *Planning algorithms.* Cambridge University Press. https://doi.org/10.1017/CBO9780511546877

Hagen Lehmann, Ingo Keller, Reza Ahmadzadeh, and Frank Broz. 2017. Naturalistic conversational gaze control for humanoid robots-a first step. *International Conference on Social Robotics.* Springer. 526–535.

Vladimir Lifschitz. 1987. On the semantics of STRIPS. In *Reasoning about Actions and Plans: Proceedings of the 1986 Workshop,* Michael P Georgeff and Amy L Lansky (Eds.), 1–9. Morgan Kaufmann Publishers.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, Matthias Grundmann. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172.*

Reuth Mirsky, Xuesu Xiao, Justin Hart, and Peter Stone. 2021. Prevention and Resolution of Conflicts in Social Navigation – a Survey. *arXiv preprint arXiv:2106.12113.*

A Jung Moon, Daniel M Troniak, Brian Gleeson, Matthew K X J Pan, Minhua Zheng, Benjamin A Blumer, Karon MacLean, and Elizabeth A Croft. 2014. Meet me where i'm gaz-

ing: how shared attention gaze affects human-robot handover timing. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction.* 334–341.

Peter Mundy and Lisa Newell. 2007. Attention, joint attention, and social cognition. *Current directions in psychological science* 16, 5, 269–274.

Jeff Pelz, Mary Hayhoe, and Russ Loeber. 2001. The coordination of eye, head, and hand movements in a natural task. *Experimental Brain Research* 139, 3, 266–277.

André Pereira, Catharine Oertel, Leonor Fermoselle, Joe Mendelson, and Joakim Gustafson. 2019. Responsive joint attention in human-robot interaction. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).* IEEE. 1080–1087.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Journal of Behavioral and Brain Sciences* 1, 4, 515–526.

Miquel Ramírez and Hector Geffner. 2009. Plan recognition as planning. *Twenty-First International Joint Conference on Artificial Intelligence.*

Miguel Ramírez and Hector Geffner. 2010. Probabilistic plan recognition using off-the-shelf classical planners. *Twenty-Fourth AAAI Conference on Artificial Intelligence.*

Michael Scaife and Jerome S Bruner. 1975. The capacity for joint visual attention in the infant. *Nature* 253, 5489, 265–266.

Ruth Schulz, Philipp Kratzer, and Marc Toussaint. 2018. Preferred interaction styles for human-robot collaboration vary over tasks with different action types. *Frontiers in Neurorobotics* 12, 36.

Shirin Sohrabi, Anton V Riabov, and Octavian Udrea. 2016. Plan Recognition as Planning Revisited. *International Joint Conference on Artificial Intelligence,* 3258–3264.

Leila Takayama, Doug Dooley, and Wendy Ju. 2011. Expressing thought: improving robot readability with animation principles. *ACM/IEEE International Conference on Human-Robot Interaction,* 69– 76.

Michael Tomasello.1995. Joint attention as social cognition. In *Joint attention: Its origins and role in development,* C Moore & P J Dunham (Eds.), (pp. 103–130). Lawrence Erlbaum Associates, Inc.

Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne, and Henrike Moll. 2005. Understanding and sharing intentions: The origins of cultural cognition. *Journal of Behavioral and Brain Sciences* 28, 5, 675–691.

Natalie Webb and Tony Renshaw. 2008. Eyetracking in HCI. In *Research Methods for Human-Computer Interaction*, P. Cairns and A. Cox (Eds.) (pp. 35-69). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511814570.004