A comparative study of Theory of Mind mechanisms and tasks between human and AI

Jan-Willem van Rhenen

4385926

12-12-23

MSc Computer Science - AI track

EEMCS

Delft University of Technology

Abstract

In order to develop artificial agents that can understand social interactions at a near-human level, it is required that these agents develop an artificial Theory of Mind; the ability to infer the mental state of others. However, developing this artificial Theory of Mind is a highly difficult process. This is because Theory of Mind is an ambiguous and multifaceted concept, having several mechanisms associated with it, and being tested using many different tasks. In this thesis, we formalize what mechanisms constitute Theory of Mind, and establish how we can represent these mechanisms using artificial intelligence. Furthermore, we evaluate whether current artificial Theory of Mind models are able to reason effectively about these mechanisms. This is done by creating Theory of Mind tasks for artificial models, evaluating their effectiveness, and allowing us to provide recommendations for the development of future artificial Theory of Mind models.

This thesis was completed as part of the Interactive Intelligence research group. The exam committee for this thesis consisted of Dr. Myrthe Tielman, Prof. Catholijn Jonker and Dr. Neil Yorke-Smith. Carolina Jorge was the daily supervisor. The writing for this thesis was completed on 5-12-23.

Contents

1 Introduction **4**

2 Related work **8**

2.1 Human Theory of Mind 8

2.1.1 Origins of Theory of Mind 8

2.1.2 Modern definition 9

2.1.3 Theory of Mind - Tasks 10

2.1.4 Theory of Mind - mechanisms 13

2.2 Artificial Theory of Mind (aToM) 15

2.2.1 Search query 16

2.2.2 Game theory-based architectures 17

2.2.3 Cognitive architectures 19

2.2.4 Observational reinforcement learning 19

2.2.5 Inverse reinforcement learning 20

2.2.6 Bayesian inference 21

2.2.7 Literature review summary 22

2.2.8 On the emergence of aToM in Large Language Models 23

3 Method **24**

3.1 Formalizing ToM mechanisms 24

3.1.1 Assumptions and definitions 24

3.1.2 Interpreting actions . . . 26
3.1.3 Visual perspective-taking . . . 29
3.1.4 Shared world knowledge . . . 30
3.2 Reconceptualizing ToM mechanisms for artificial intelligence . . . 33
3.2.1 On rule-based architectures . . . 33
3.2.2 Visual perspective-taking . . . 34
3.2.3 Interpreting actions . . . 36
3.2.4 Shared world knowledge . . . 38
3.2.5 Summary . . . 40
3.3 Experiment design . . . 41
3.3.1 Base components . . . 41
3.3.2 Control experiment . . . 42
3.3.3 Visual perspective-taking experiment . . . 42
3.3.4 Interpreting actions experiment . . . 43
3.3.5 Learning knowledge experiment . . . 44
3.3.6 Desire and intentions experiment . . . 45

4 Use case **46**
4.1 Model specification . . . 46
4.2 Results . . . 47

5 Conclusion and Discussion **48**
5.1 RQ1.1: Relevant aToM mechanisms . . . 48
5.2 RQ1.2: Requirements for aToM models . . . 50

5.3 RQ1.3: Evaluating aToM models 51
5.4 Future work 52
5.5 Conclusion 52

Chapter 1 - Introduction

Theory of Mind (ToM) is an important cognitive ability that humans use to infer the mental states of others [17]. ToM is an essential component in human social interaction, allowing humans to understand and react to the intentions, beliefs, desires, and emotions that underlie the behavior of the ones they're interacting with [8]. Currently, artificial intelligence (AI) does not possess ToM skills at a human level [7]. If AI is to collaborate with humans in social situations to a level akin to how humans interact amongst themselves, a developed artificial Theory of Mind (aToM) is a prerequisite [94]. However, ToM is a complex and multifaceted ability, encompassing different mechanisms and being measured using various tasks. It is currently unknown how we can best translate these mechanisms and tasks in an artificial context. This thesis will research these mechanisms and tasks, looking at how we can formalize and evaluate the different mechanisms relating to (artificial) Theory of Mind.

Over the past couple of years, the field of AI has rapidly advanced to such a degree, that human-level performance is now achievable on a wide range of tasks. For example, deep learning models can now recognize the contents of images at an accuracy similar to humans [66]. Similarly, large language models, such as GPT-4 by OpenAI [64], now outperform humans in language processing tasks to such an extent, that many sectors and legislative bodies have already made agreements to limit the use of artificial intelligence in order to prevent job loss [78, 2]. Despite these agreements, agents possessing AI continue to be implemented across many industries [58]. As impressive as these achievements might be, currently AI is still a tool for humans to use. True virtual assistants, that can autonomously reason and interact with others at a human level, do not yet exist [94].

If AI is to serve as an assistant to humankind, it needs to be able to understand social interactions at a human level. This is necessary, so that this AI can understand what their human counterpart wants and needs, but likewise, is also necessary so that humans can understand AI. If humans can understand AI decision-making better, they will show higher levels of trust towards these systems [4]. With higher levels of trust, humans will be more likely to delegate tasks to their virtual assistants. On the contrary, if people do not have faith in their artificial counterparts, they are unlikely to rely on decisions made by these agents, even if those decisions are correct [68]. Therefore, if AI is to serve as a reliable partner for humankind, a high degree of social cognition is required in order to promote human-AI relations [94].

As stated before, one of the core abilities that lead to this social cognition is called Theory of Mind (ToM) [94]. Possessing ToM is strongly linked with both trustworthiness [84], as well as the ability to make accurate judgments relating to trust [88]. ToM is broadly defined as the ability of an individual to impute the mental state of others [72]. Using ToM, humans can reason about beliefs, knowledge, desires and intentions of others through observation [6]. ToM is a multifaceted concept that encompasses many different mechanisms. Humans learn these mechanisms as they develop into adulthood. For instance, at four years old, most children are able to understand that a person can hold a false belief, and can reason about what it means for this person to hold a false belief [12]. However, passing a 'false belief' task, does not necessarily mean that a child has then successfully

acquired ToM. Although a four year old can reason about false beliefs, they usually can not yet understand the concept of another person pretending to believe or feel some way, even though they believe or feel something else [90, 92]. Understanding that someone might pretend to believe in something is also an important mechanism to master for the acquisition of ToM.

Although ToM is broadly defined as the ability of an individual to impute the mental state of others [72], a single, more precise definition, that defines what mechanisms and tasks are part of this concept, is notably absent. This is, because ToM is studied in a wide range of fields, such as developmental psychology, ethology and psychiatry [6]. These fields study specific facets of ToM, using many different types of tests [12, 38, 14, 56]. Given the wide range of disciplines involved, there's a notable lack of uniformity in defining ToM, with definitions frequently becoming confounded by adjacent terms. For instance, Quesque and Rosseti [76] found that definitions for ToM sometimes overlap with definitions for: "mentalizing", "mindreading", "perspective-taking", "empathy", "cognitive empathy", or "empathic perspective-taking". However, in other cases these terms can refer to entirely different concepts, e.g. Cuff et al. [20] found 43 different definitions for the term 'empathy'. This heterogeneity and nonspecificity of definitions is also present in the types of tasks used to test ToM (or any adjacent concepts like the ones mentioned above - see Figure 1.1) [76]. This ambiguity surrounding ToM and ToM testing makes it a very complex concept to research. A measure for 'true' ToM does not exist, and ToM can only be tested for indirectly through the use of tasks [17].

Although human ToM has been a topic of research for over 40 years [95], artificial ToM research is still in its early stages. Initial attempts have been made at creating artificial ToM. Several papers [96, 77, 74, 62] exist that have created a model for solving tasks relating to ToM. Most papers use some architecture based on deep learning, reinforcement learning, or a combination thereof [94, 21, 7]. However, artificial ToM (aToM) research unfortunately suffers from similar ambiguity issues as its human counterpart. The equivocal nature of ToM makes it easy to relate many different fields and topics to this concept. Many current papers choose to focus on a specific facet of ToM that is suitable for their purposes. As a consequence, it becomes very difficult to compare different papers at face value. Papers that outwardly appear to be similar (the creation of artificial ToM-like qualities using machine learning), can, upon closer inspection, be quite different in scope and claims. Different research, for instance, focuses on replicating certain mechanisms relating to human ToM [96], attempts to develop an AI that is human-interpretable [62], or combines ToM with related AI fields like opponent modeling [77].

Besides issues relating to the ambiguity of ToM as a concept, artificial ToM research also suffers from additional difficulties, resulting from the translation of ToM to an artificial context. In their paper: "Mind the gap: Challenges of deep learning approaches to Theory of Mind", Aru et al. [7] mention several key issues. 1) The grand goal of aToM research is to create a model that has human level ToM. However, what that goal entails is not precisely defined. As stated before, ToM is an ambiguous and multifaceted concept, and humans achieve ToM through a variety of mechanisms. What mechanisms should an aToM model learn in order to achieve human level ToM? 2) In order to possess ToM, one has to be able to reason about abstract concepts related to the mind. These abstract concepts can't be measured directly. In order to evaluate ToM skills in humans, researchers designed tasks that allow us to indirectly measure the level of ToM of a participant [76, 17]. However, these tasks are designed for human participants. In order to apply them in an artificial environment,

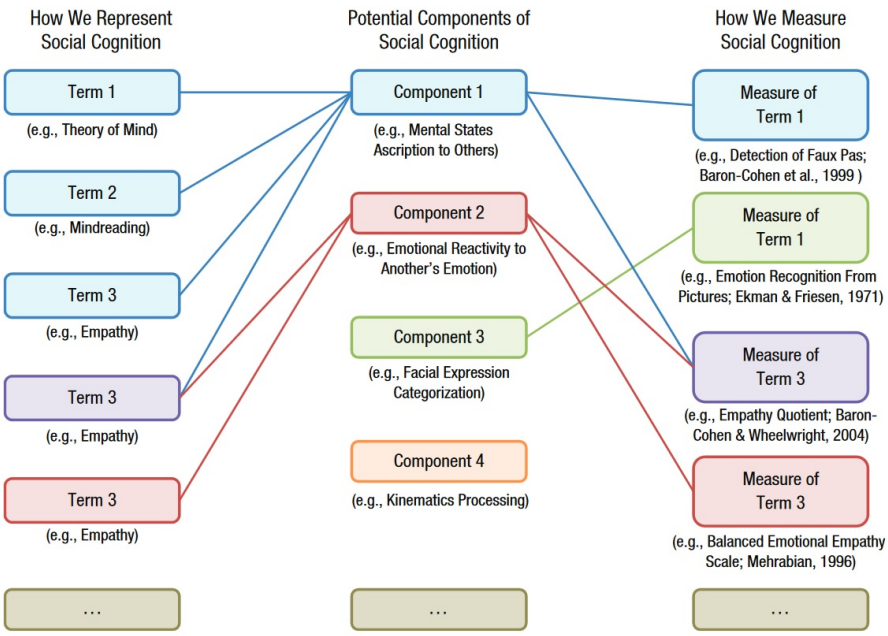

Figure 1.1: Schematic depiction of the heterogeneity and nonspecificity of both representation and testing of concepts relating to social cognition. Each color in the depiction represents one component. Different papers can refer to one component by using a different term (e.g. ToM, mindreading and empathy all refer to ascribing mental states to others). This is what is meant by heterogeneity. At other times, multiple papers use the same term to refer to different concepts (e.g. three papers mean something different when referring to empathy). This is what's meant by nonspecifity. Both heterogeneity and nonspecifity also apply to tasks used to test these components. Figure taken from [76].

some translation will have to be made in terms of format. Aru et al. [7] found little research in this direction. Most papers designing an aToM model attempt to design tasks that are outwardly similar to human ToM tasks [96, 77, 74, 62]. However, because machine learning algorithms solve problems in a different manner than how humans solve problems (another key problem mentioned by Aru et al.), this might not produce a model that possesses true ToM. This problem is highly related to the following problem: 3) Although machine learning algorithms can be applied to a variety of problems, they are still designed in such a way that allow them to perform well at some types of problems, but worse at some others. This fact is well known, and is summarized in the 'no free lunch' theorem [97]. Currently, several different types of machine learning algorithms are researched in aToM research [33]. Although their general strengths and weaknesses are known, little research exists on how these general strengths and weaknesses relate to ToM mechanisms and tasks. What is required of machine learning models in order to be able to reason about ToM mechanisms and tasks?

Considering these difficulties, the creation of true artificial Theory of Mind has been lauded as one of the Grand Challenges of Science Robotics [99]. More research is required on each of these

aforementioned problems in order to eventually develop this artificial ToM. Thus, the goal of this thesis is to address each of these problems. We will analyze what mechanisms underlying ToM are relevant to aToM research, as well as how we can evaluate aToM models on these mechanisms. In order to do so, we will investigate mechanisms and tasks relevant to human ToM research, and compare whether, and how, these mechanisms and tasks are researched in an artificial context. Frameworks for how humans reason about these relevant mechanisms will then be provided. This addresses the first problem: What mechanisms should an aToM model learn in order to achieve ToM? Using these human frameworks as a basis, we can reason about what is required of aToM models to successfully solve tasks relating to these mechanisms. We will discuss these requirements, and see how we can accommodate for them when designing aToM tasks. This addresses the third problem mentioned; what is required models to reason effectively about ToM. Finally, we will discuss how to design aToM testing tasks, providing sample tasks and evaluating a single state-of-the-art aToM model on it as a proof-of-concept, presenting sample solutions to the second problem. This leads to the following research questions:

RQ1: How can one formally define the different mechanisms relating to ToM, and evaluate aToM models' effectiveness in reasoning about said mechanisms?

RQ1.1: What mechanisms are relevant for state-of-the-art aToM research?

RQ1.2: What are the requirements for aToM models to be able to effectively reason about these mechanisms?

RQ1.3: How can these mechanisms be evaluated for artificial ToM models?

In order to answer these research questions, we will establish what tasks are currently used for ToM research, as well as which mechanisms are required in order to solve these tasks. Furthermore, we will analyze the most commonly used model architectures for aToM research. These evaluations of tasks, mechanisms and model architectures can be found in chapter 2. In chapter 3 we will then formalize the mechanisms found in chapter 2, creating a framework for how humans reason about each mechanism. These frameworks are then reconceptualized in an artificial context, allowing us to reason about what is required of aToM models in order to effectively reason about mechanisms related to ToM. Furthermore, we will use these frameworks to design tasks to evaluate aToM models' effectiveness at reasoning about these mechanisms. These tasks are applied to an existing aToM model in chapter 4, as a proof-of-concept. A discussion of our findings, as well as recommendations for future work can be found in chapter 5.

Chapter 2 - Related work

2.1 Human Theory of Mind

This section will outline relevant literature on human ToM. The goal of this section is to provide a clear definition of human ToM, as well as focus on what mechanisms humans employ for ToM. In finding these mechanisms, we will take a task-centric approach. This section addresses the following: What tasks are available in human ToM research, and can we group these tasks based on what mechanisms are required in order to reason about them? In order to find these tasks and mechanisms, we will first analyze ToM as a concept, looking at the history and more contemporary research, in order to provide the reader with a modern definition of ToM.

2.1.1 Origins of Theory of Mind

The first mention of the term 'Theory of Mind' is in a paper from 1978 by Premack and Woodruff [72]. In this paper, they show that chimpanzees possess a Theory of Mind, which is similar to our own. In their paper, Premack and Woodruff provide the first definition of the term 'Theory of Mind': "The individual imputes mental states to himself and to others" (page 515). This original definition was rather concise, and therefore benefited from additional explanation. In a commentary on their paper, Pylyshyn [73] explicated the definition in order to make it more precise. Pylyshyn gave the following definition of ToM: An individual having a representation of a state of affairs, a relationship to this state (desiring x, needing x, thinking about x), and most of all having an explicit definition of this relationship. This explicit definition is referred to as an ability for 'meta-representation' by an individual (page 593). This definition can best be explained through an example story, for instance: a boy is thirsty. He wants to drink water. He also knows that he wants to drink water, because it will alleviate his thirst. In this example, the boy has a representation of a state of affairs (drinking water), a relationship to this state (wanting to drink water) and a representation of this relation (he knows what it is like to 'want to drink water'). He can, for instance, think of circumstances under which he would like to drink water (e.g. when he is thirsty). The fact that this boy can represent 'wanting to drink water', means he can not only apply it to himself, but also to others (e.g. "my friend tells me he is thirsty. He must want to drink water.").

After the term 'Theory of Mind' was first used by Premack and Woodruff [72], the term gained traction in the field of developmental psychology when H. Wimmer and J. Perner wrote their 1983 paper [95] on young children's representation and understanding of deception and/or wrong beliefs. In this paper, they designed an experiment in which children listened to a story about a boy named Maxi, placing a chocolate cake in one of three (in case of one version of the story) or two (in case of another version of the story) different colored cupboards and then leaving the scene. The chocolate cake is then moved to a different cupboard by his mother. When Maxi comes back the story diverges

in two separate versions. One version has him telling his grandfather where the cake is, so they can share it (so Maxi will be truthful about where the cake is). The other version has Maxi's big brother asking where the cake is, so he can eat it. The children listen to Maxi thinking to himself that he will lie to his brother, so that his brother does not eat the cake [95]. The children are then asked in which cupboard Maxi will say the cake is. This allows them to demonstrate an understanding of ToM [95].

This experiment was later simplified by S. Baron-Cohen, A.M. Leslie and U. Frith, in their research of ToM in autistic children [12]. This was done by removing the part of the story in which Maxi lies or tells the truth (and it is assumed that Maxi always tells the truth). The names of the children were changed to Sally and Anne, and the Sally-Anne experiment is considered a benchmark task for the development of ToM in children [51, 91].

2.1.2 Modern definition

Over the years, ToM has transitioned from its origins in behavioral sciences [72] and developmental psychology [95, 12], to become a subject of extensive research in a diverse range of disciplines, such as cognition, philosophy and ethology [94]. Each of these fields considers ToM from a different perspective and, more importantly, uses different tasks to measure mechanisms related to ToM [76]. This wide academic pursuit of ToM has made the concept highly ambiguous and conflated with adjacent concepts (such as 'empathy', 'mentalizing' or 'mindreading'), as well as tasks to test said adjacent concepts (for a schematic depiction of this, see Figure 1.1). Quesque and Rossetti [76] did a meta-analysis of the most commonly used tasks in order to test for ToM in humans. We will discuss the tasks they studied below (See subsection 2.1.3). In their analysis, Quesque and Rossetti took a critical look at what tasks measure ToM, and what tasks measure adjacent mechanisms. In order to do so, a more thorough view of ToM as a concept was required.

ToM is viewed as inferring the mental states of others [72]. This mental state includes several components, such as belief, intention and emotional inferences [30], but several studies [27, 26, 45, 37] have given evidence to validate that ToM also encompasses the ability to infer how another represents the surrounding world. How humans achieve this ability is a lengthy and complex topic of its own, so it will not be discussed here in detail. However, it is important to note that a variety of social cognitive subcomponents are employed in order to achieve ToM in a variety of circumstances [31, 32, 79]. This supports the view that ToM can be viewed as a singular process that relies on a diverse set of lower-level mechanisms.

Using the view that ToM is a process consisting of multiple mechanisms, with the mechanism of inferring the mental state of others at its core, Quesque and Rosetti [76] formulated two criteria that are required of ToM-testing tasks:

1. **Nonmerging criterion:** Participants need to make a distinction between their own mental state, and the mental state they infer. Attributing a mental state to someone else that is similar to their own does not constitute ToM (e.g. I can see a ball, now I infer that the person standing right besides me also sees a ball. This is not considered ToM).

2. **Mentalizing criterion:** Lower-level processes should not account for successful performance on ToM tasks. For instance, emotion recognition is an often cited ToM task [17, 14]. However, frequently these tasks can also be explained as visual discrimination, e.g. we can distinguish a happy face from a sad face, just like we can distinguish an appple from a pear. No mentalizing of the other's state-of-mind is necessary to make this distinction. Only if a participant is asked whether the person is actually happy (or sad), they would need to engage in ToM, as visual discrimination is an inadequate mechanism to answer that question.

The nonmerging criterion ensures that only tasks that ask a participant to consider a problem from different perspectives are included as ToM tasks. This criterion is included, because there is supporting evidence that this ability to co-represent multiple perspectives is key to all perspective taking processes involved in social scenarios [25, 27]. The mentalizing criterion exists to ensure that inferring the mental states of others is essential to solving the task, as tasks that fulfill this criterion can not be explained by lower-level mechanisms alone. This criterion is in essence an application of Occam's Razor. Although Occam's Razor as a principle is not infallible in science [22], this criterion shows that we can't conclusively prove that humans employ ToM in order to solve tasks that do not satisfy it.

2.1.3 Theory of Mind - Tasks

Using the nonmerging and mentalizing criteria, Quesque and Rossetti [76] analyzed 22 of the most commonly cited ToM tasks. A full overview of the tasks can be found in their paper. However, in this subsection only the tasks that fit both the nonmerging and mentalizing criteria are discussed, as according to Quesque and Rossetti, only these tasks can be used to measure ToM. Only eight tasks were found that fit both criteria. The tasks are given a name for easy reference.

False belief task The Sally-Anne task, mentioned in subsection 2.1.1, is one of the tasks that passes both criteria. Scenarios like the Sally-Anne task are called 'false-belief' tasks, and have been researched extensively [95, 12, 28]. These false-belief tasks ask that a participant reasons about someone else's perspective, that is different from their own (nonmerging criterion) and almost exclusively asks that a participant mentalizes someone's state-of-mind (mentalizing criterion), besides some listening comprehension skills and verbal skills. Listening comprehension and verbal skills are a prerequisite for all tasks mentioned, and it seems that these are an unavoidable prerequisite for solving ToM tasks.

Faux-pas task Another, closely related tasks that fits both criteria is about the detection of so-called 'faux-pas' scenarios [13]. In these scenarios, which are being told as a story to the participant, some actor does or says something that is socially frowned upon, or in other words, commits a social faux-pas. For example, take a look at the following story from the original paper:

Kim helped her Mum make an apple pie for her uncle when he came to visit. She carried it out of

the kitchen. "I made it just for you," said Kim. "Mmm", replied Uncle Tom, "That looks lovely. I love pies, except for apple, of course!"

The participant is then asked whether someone in this story said something they shouldn't have said. This task also tests false belief skills of the participant ("Did Uncle Tom know that the pie was an apple pie?") and fits both the nonmerging and mentalizing criteria, like the false belief task.

Burden of knowledge task Tasks testing so-called burden of knowledge are another type of task that fits both nonmerging and mentalizing criteria [46] (also see [47]). In the task, the participant is presented with an actor (Mark) giving a statement to another actor (June). This statement could be interpreted as sarcastic, or could be sincere. The participant has to give the likelihood that June believes that Mark's statement is sarcastic. However, participants are presented with additional information, so that they know whether Mark's statement was sarcastic or not. Therefore they have to separate what they know from what June knows in order to answer the question. A majority of adults do not achieve high accuracy on burden of knowledge tasks.

Figure 2.1: The spatial orientation task. In this specific task, participants are asked the following: *'Imagine you are at the stop sign and facing the house. Point at the traffic light.'* Participants have to separate reality from the situation presented in the picture, and mentalize themselves to be in a simulated world. Task taken from [40].

Spatial orientation task The fourth type of task that passes both criteria is a spatial orientation task [40]. Participants are presented with a bird's eye view of a scene, and have to imagine themselves inside the scene, in order to provide the location of objects, relative to their position (see Figure 2.1). Interestingly, in this task, participants are not asked to reason about other humans, but have to simulate an environment instead. This task is included as a ToM task, as in their definition of ToM, Quesque and Rossetti [76] included the large body of evidence [27, 26, 45, 37] that supports

ToM as including: someone inferring how others represent the environment (see subsection 2.1.2). Furthermore, this task passes both criteria. In the task, participants have to separate their own mental state from the (different) mental state of their simulated selves in the scene. In addition, it is required that the participant mentalizes the simulated environment. It could be argued that spatial orientation skills are also necessary for this type of task. However, mentalizing is still a necessary component.

Perspective-taking task A very similar task to the spatial orientation task exists. In this task, the participant is asked to represent or describe how a scene would be viewed by another actor, that is in a different location in the same scene. This task, developed by Piaget and Inhelder [71] was originally developed in 1956 and, thus, has existed before the advent of the term 'Theory of Mind'. It however fits both criteria in a similar fashion to the spatial orientation task mentioned in the previous paragraph.

Director's task The director's task [98] is another spatial task with perspective-taking. The participant is shown an open cabinet containing an assortment of objects. Some objects are duplicate (e.g. 2 apples). Some objects can be viewed from both the front and the back of the cabinet. However, other objects are hidden from the backside by means of a back panel. Then, the participant is asked to move certain objects by an actor that is standing behind the cabinet. In instances where there are duplicate objects, one of which is hidden from the backside, the participant has to infer the perspective of the actor, in order to deduce which object should be moved. Again, this task fits both the nonmerging and mentalizing criteria.

Strange stories task A different type of task, testing the participant's ToM skills related to social situations is called the strange stories task [38]. In this task, the participant has to provide an explanation for the mental state of an actor, by carefully regarding the surrounding context. For instance, consider the following story:

Katie and Emma are playing in the house. Emma picks up a banana from the fruit bowl and holds it up to her ear. She says to Katie, "Look! This banana is a telephone!"

In order to explain Emma's actions, one has to understand that Katie and Emma are engaging in pretend-play. Therefore this type of test requires a degree of social awareness in addition to the mentalizing of Emma's beliefs and intentions. It is interesting to note that Emma's beliefs, in this specific example, are not different from the participant's own beliefs (i.e. the banana is not a telephone). One could therefore posit that the nonmerging criterion does not apply in this instance. However, Emma's intentions do differ from the participant's own intentions. She wants to pretend that the banana is a telephone, whereas the participant (supposedly) does not want to do this. Intention is included in Quesque and Rossetti's definition of ToM [76] (see subsection 2.1.2). Therefore, this type of task satisfies the nonmerging criterion.

MASC task The last ToM task Quesque and Rossetti [76] found, is called: Movie for the Assessment of Social Cognition (MASC) [23]. In this task, participants are shown a 15 minute movie, and have to answer questions relating to the emotions, thoughts and intentions of the characters in the movie. This task is highly similar to the strange stories task [38], but combines this with several tasks relating to emotional inference, such as the 'Reading the Mind in the Eyes Test' (RMET) [14]. These type of emotional inference tasks are traditionally associated with ToM, but Quesque and Rossetti [76] found that these types of tasks fail on both the nonmerging and mentalizing criteria. This is, because in order to recognize an emotion on someone's face, no mentalizing of a (different) mental state is required. One can simply recognize said emotion, without understanding the underlying beliefs and intentions that led to that emotion. Therefore emotional inference tasks are not included as ToM tasks by Quesque and Rossetti [76]. The MASC task, however, is included, based on the merit that, besides emotional inference questions, the task also includes questions on thoughts and intentions.

2.1.4 Theory of Mind - mechanisms

Now that a clear overview of the tasks used to test ToM is presented, we can categorize these tasks based on what mechanisms humans use in order to solve them. This is done in order to find what mechanisms underly ToM in humans. Byom and Mutlu [17] did a meta-analysis of task types used for ToM testing, in order to identify these mechanisms. They found three categories of tasks. Each task in the same category requires a similar mechanism in order to solve them. They named these categories based on the required mechanism: Perceiving social cues, interpreting actions and shared world knowledge. However, in their study they considered different tasks than Quesque and Rossetti [76] (for an overview, see Table 2.1). In this subsection, we will compare the tasks considered by Byom and Mutlu [17], to the tasks found by Quesque and Rossetti [76]. Using this comparison, we categorize the tasks found by Quesque and Rossetti in terms of mechanisms required to solve them. For this categorization, we will use the study by Byom and Mutlu as a basis, expanding upon it where necessary. We will start by discussing the mechanisms Byom and Mutlu found.

Perceiving social cues We will start by discussing the mechanism of 'perceiving social cues'. For this mechanism, Byom and Mutlu considered two tasks: the 'Reading the Mind in the Eyes Task' (RMET) [14] and the 'The Awareness of Social Inference Task' (TASIT) [56]. Of these two tasks, the RMET task is also considered by Quesque and Rossetti. They also consider several other so-called 'emotional inference/emotion recognition' tasks. However, they find that none of these emotional inference tasks pass the nonmerging and mentalizing criteria (see subsection 2.1.2). Therefore, we can posit that the mechanism of 'perceiving social cues' is not a mechanism related to ToM.

Interpreting actions The second mechanism Byom and Mutlu consider is called 'interpreting actions'. Research on the development of ToM has shown that 6 month old babies already have some rudimentary expectations regarding how humans interact with other humans and inanimate objects [54]. Humans generally hold the belief that others act in a manner consistent with their beliefs and goals [41, 3]. Using this assumption of consistency, one can make inferences about the

beliefs and goals of another by observing their behavior.

In order to arrive at this mechanism of 'interpreting actions', Byom and Mutlu exclusively considered false belief tasks [95, 28, 70] (see subsection 2.1.1). It can therefore be assumed that Byom and Mutlu believe 'interpreting actions' is the mechanism humans use to infer false beliefs. Conversely, it can be assumed that 'interpreting actions' is not a mechanism that is used for other types of ToM-related inferences. Quesque and Rossetti consider one of these false belief tasks [95] (see subsection 2.1.3 - false belief task). However, they also consider several other tasks that include a false belief component [13] (see subsection 2.1.3 - faux-pas task) or require the separation of the participant's beliefs from that of the actor [46] (see subsection 2.1.3 - burden of knowledge task).

Shared world knowledge The final mechanism Byom and Mutlu consider is 'shared world knowledge'. According to them, Shared world knowledge is a skill that is for instance tested in humans during conversation. Conversation requires participants to make use of cues from the conversational partners, as well as any previously learned knowledge about the world. For instance, suppose you have a first date with someone, and they mention that their favorite animal is a goat. You, therefore, take them on a date to the petting zoo, because you can relate that someone liking goats means they would enjoy seeing them. You also know that goats are animals, and that those animals are regularly kept at petting zoos.

For this mechanism, Byom and Mutlu consider the strange stories task [38], a task also considered by Quesque and Rossetti (see subsection 2.1.3 - strange stories task). Byom and Mutlu also consider a task called the 'character intention task' [80], a task not considered by Quesque and Rossetti: This task tests the ability of participants to correctly infer the intention of a character, by choosing the last frame of a comic strip from several possible panels. One of the biggest advantages of this way of testing is that it does not rely on text or audio in order to be completed, but can be performed on a purely visual basis. This helps to test for ToM in persons that have problems with word understanding and/or processing.

Thus far we have categorized four out of eight tasks considered by Quesque and Rossetti (see subsection 2.1.3) in the mechanisms found by Byom and Mutlu: The false belief task, the faux-pas task (partially) and the burden of knowledge task can be considered to require 'interpreting actions'. The strange stories task can be considered to require 'shared world knowledge'. This leaves the spatial orientation task, the perspective-taking task, the director's task and the MASC task. The MASC task (see subsection 2.1.3 - MASC task) states that it is based on the strange stories task [38], combined with emotion recognition tasks like the RMET task [14]. As stated before, emotion recognition tasks do not test ToM-related mechanisms, so it can be derived that the ToM mechanism required for the MASC task is 'shared world knowledge'.

Visual perspective-taking Thus three tasks are left: The spatial orientation task, the perspective-taking task and the director's task (see subsection 2.1.3). These three tasks are all related to perspective taking in a physical sense, i.e. inferring what another person can observe from their point of view. These tasks were not considered by Byom and Mutlu, because they do not align with a historical definition of ToM (see subsection 2.1.1). However, Quesque and Rossetti note several studies

that present evidence that ToM also encompasses the ability to infer how another represents the surrounding world [27, 26, 45, 37]. Because of these studies, as well as these three tasks passing the nonmerging and mentalizing criteria, Quesque and Rossetti include them in their study. These three tasks are all highly related to eachother, and do not require a participant to interpret the actions of another person, nor do they rely on an understanding of previously learned knowledge. Therefore we propose to categorize these three tasks as requiring another mechanism: 'visual perspective-taking'.[1]

Mechanism	**Tasks by Quesque and Rosetti [76]**	**Tasks by Byom and Mutlu [17]**	**Tasks by both**
Interpreting actions	Keysar (1994) [46]	Flavell et al. (1983) [28] Perner and Wimmer (1985) [70]	Wimmer and Perner (1983) [95]
Shared world knowledge	Dziobek et al. (2006) [23] Piaget and Inhelder (1956) [71]	Sarfati et al. (1997) [80]	Happé (1994) [38]
Visual perspective-taking	Hegarty and Waller (2004) [40] Wu and Keysar (2007) [98]		
Perceiving social cues*	Heider and Simmel (1944) [41]	McDonald et al. (2006) [56]	Baron-Cohen et al. (2001) [14]

Table 2.1: Overview of the tasks analyzed by Quesque and Rosetti [76] and/or Byom and Mutlu [17]. Tasks are divided per category, based on what mechanism is required in order to solve a task. The categories of 'interpreting actions' and 'shared world knowledge' were defined by Byom and Mutlu. Tasks in the category of 'visual perspective-taking' were only considered by Quesque and Rossetti and categorized by this author, based on the work by Byom and Mutlu. Tasks categorized in the mechanism of 'perceiving social cues' (marked with an asterisk here) were not considered ToM tasks by Quesque and Rossetti according to their criteria (see subsection 2.1.2).

Summary So far, we have derived a modern definition of ToM (see subsection 2.1.2) and two criteria for what is required of tasks to test ToM skills. Using these two criteria we found eight tasks commonly used to test ToM skills in humans. These tasks were grouped into three categories, based on what mechanism humans employ to solve these them. For these mechanism-based categories we used the research by Byom and Mutlu [17] as a basis, but expanded upon it by adding a new mechanism: 'visual perspective-taking'. It is important to note that the mechanisms presented by Byom and Mutlu are not precise, and are mostly a broad grouping of tasks. In a later chapter we will attempt to formulate what these mechanisms precisely entail, based on the tasks present in each category.

2.2 Artificial Theory of Mind (aToM)

Now we have established what tasks and mechanisms are present in human ToM research, we can review aToM research. The goal of this review is to find what architectures are commonly used for aToM research. In order to find what is required for aToM models to reason about ToM mechanisms, and design methods to evaluate aToM models on these mechanisms (answering *RQ1.2* and *RQ1.3* - see chapter 1), it is necessary to know what types of architectures are the most commonly used in aToM research. We will look at three factors: What model architectures are present in aToM

[1] It is important to note that, while all of these tasks, and the research done by [27, 26, 45, 37] focuses purely on perspective-taking through a visual medium, Quesque and Rossetti [76] believe that this perspective-taking also applies to other sensory modalities.

research, what types of tasks they use to test their models, and how well do these models generalize. For this research, only models that generalize well will be taken into account. After all, if a certain model does not generalize, and is designed for a specific problem, it can not be evaluated for anything other than that problem itself. This makes these types of architectures unsuitable for developing aToM at a (near-)human level. This does not mean that papers developing these types of models do not contribute to aToM research, but rather that they are unsuitable for the research performed in this thesis.

In order to do this review, we looked at 15 papers (see Table 2.2), found through a query on Elsevier's scopus [1] (see subsection 2.2.1). This research was then combined with two review papers: Gonzalez and Chang [33], and Langley et al. [53]. Both of these papers are focused on categorizing each architecture present in aToM research, highlighting their strengths and weaknesses. As stated in the introduction of this thesis, each type of architecture used has defined strengths and weaknesses (see the 'no free lunch' theorem [97]). These strengths and weaknesses make certain architectures suited towards certain types of ToM tasks. We will discuss these strengths and weaknesses. We will group the papers by the type of architecture used (as found by both review papers [33, 53]). The architectures discussed are: Game theory-based architectures, cognitive architectures, observational reinforcement learning, inverse reinforcement learning and Bayesian inference. It is important to note that two of these architecture categories are rule-based (game theory-based and cognitive), i.e. fully designed by humans. The other three types of architectures, although partly designed by humans, contain a learning component, i.e. machine learning algorithms.

2.2.1 Search query

A literature review was done on 15 papers (see Table 2.2) that designed and/or implemented artificial ToM models. These papers were selected through a search query on Elsevier's scopus [1]. In order to find relevant papers focusing on researching ToM in an artificial context, the search query was constructed as follows: A title including 'Theory of Mind' and/or 'ToM', and the words 'artificial' or 'AI', and 'model' in the title, abstract and/or keywords. The word 'artificial' or 'AI' were included to exclude papers researching human ToM. The word 'model' was included to filter out papers that didn't design an aToM model, but focused on other types of aToM research. To further refine the search results, only papers from computer science were taken into account.[2] For relevancy, papers were sorted by number of citations, and only papers with 5 or more citations were taken into account. This query found 23 papers. During review of individual papers, 8 papers were found to not be suitable for the review. Therefore they were excluded. The excluded papers were as follows:

- **Cominelli et al. [19]:** This paper is mostly focused on robotics, and the recognition and displaying of emotions. ToM is mentioned in the paper, but only constitutes a small part of

[2]One can use the following search query to find these results on Scopus [1]: *(TITLE (theory AND of AND mind OR tom) AND TITLE-ABS-KEY (artificial OR ai AND model)) AND PUBYEAR > 1999 AND PUBYEAR < 2023 AND (LIMIT-TO (SUBJAREA , "COMP")).* The publication years were included in this query for reproducibility, because all papers found at the time of writing (7-11-2023) were published between 2000 and 2022. However, Scopus does not allow one to search on number of citations at a certain date, so more papers with 5 or more citations might be found if this query is used at a later date.

the total paper.

- **Ono et al. [63]:** Created a model, so that humans can better understand synthetic utterances made by a robot, using ToM. The robot in their study does not have any autonomous decisionmaking.
- **Friedlander and Franklin [29]:** Presents a conceptual model of cognition, but provided no implementation.
- **Eliasmith [24]:** Gives an argument against dynamicism providing a convincing alternative to currently available cognitive theories. This paper is purely about human psychology.
- **Bara et al. [11]:** Provides a dataset of collaborative tasks that can be performed by pairs of two humans in a virtual environment.
- **Williams et al. [94]:** Discusses the role of aToM in artificial social intelligence. Does not include a model.
- **Melhart et al. [57]:** Researches player preferences using Support Vector Machine-based [61] preference learning. Does not include an aToM model.
- **Langley et al. [53]:** This paper is a review of the architectures and approaches currently used in aToM research. This paper will be used to guide this literature review, together with Gonzalez and Chang [33]. However, because it does not contain a model itself, it will not be analyzed like the other 18 papers that do include a model.

The subsections below will discuss each category of architecture, as found by Gonzalez and Chang [33] and Langley et al. [53]. For each category of architectures, we will discuss common goals and tasks, and hypothesize how their strengths and weaknesses relate to specific ToM mechanisms (as defined in subsection 2.1.4).

Game theory	**Cognitive**	**Observational RL**	**Inverse RL**	**Bayesian inference**
Hiatt et al. [42]	Bosse et al. [15]	Oguntola et al. [62]	Winfield et al. [96]	*Baker et al.** [10]
Stühlmuller and Goodman [83]	Sarkadi et al. [81]	Nguyen and Gonzalez [60]	Rabinowitz et al. [77]	Patacchiola et al. [69]
Klatt et al. [49]	Sarkadi et al. [82]			
Pynadath et al. [75]	Panisson et al. [67]			
Veltman et al. [87]	Vossen et al. [89]			

Table 2.2: Overview of the papers reviewed in this section. Papers are categorized based on the type of architecture used by the model they describe. An explanation of each of these categories can be found in the subsections below. It should be noted that Baker et al.'s paper [10] (marked with a *) was not found during the literature review, but rather was mentioned in both review papers [33, 53] as well as in several of the other papers found during the literature review [60, 62]. Therefore it is included.

2.2.2 Game theory-based architectures

Some of the earlier developed models are based on game theory [33]. Game theory is a branch of economics that studies interactions between rational decisionmakers [65]. Models based on game

theory are rule-based, following the principles outlined in game theory. Game theory-inspired models operate on the assumption that each actor in a scene will want to maximize their reward. In order to maximize their reward, an actor will make strategic decisions, taking into account what decisions other actors might make. This 'taking into account of other's decisions' is what is attributed to possessing ToM. Some of these game theory models even take into account that other actors have their own ToM, and incorporate that knowledge into their own decision-making. This would correspond to a second-order ToM (e.g. I believe that he believes that I believe).

Game theory-based models operate on the assumption that an actor attempts to maximize their personal reward. Because these models operate on this specific assumption, they are unable to reason about an actor's intention. After all, the only possible intention for an actor is always assumed to be to maximize their personal reward. This would mean that they are unable to reason about problems that require the participant to reason about intention, such as the strange stories task or the MASC task (see subsection 2.1.3). A more modern variant of game theory-based models, Psychology Game Theory (PGT) models, attempts to solve this shortcoming by trying to capture beliefs, intentions and emotional inferences in a singular utility function. It is then assumed that each actor attempts to maximize their own utility function [33]. However, this makes PGT models highly dependent on the choice of utility function. Therefore it can be hypothesized that game theory-based models are weak at reasoning about the mechanism of 'shared world knowledge'. The other two mechanisms also require some modifications in the formulation of the task. For instance, in the Sally-Anne task, the utility/reward function would have to be tuned in such a way that Sally is rewarded when she moves to where she believes the marble is.

In the literature review we find five papers that use game theory-based models. Hiatt et al. [42] use a probabilistic model to consider different hypotheses to explain the behavior of their human partner in a team based scenario. They note that their model does not scale well beyond simple problems with a limited number of hypotheses. Stühlmuller and Goodman [83] use a probabilistic model with nested conditions to make decisions in games, such as tic-tac-toe. Their models are highly specific to each game they test, i.e. they have a different model for each game, and therefore do not generalize. Klatt et al. [49] use a similarly narrow model (based on an architecture named Psychsim), in order to model safe-sex negotiations in the context of aids prevention. Pynadath et al. [75] uses the same architecture as Klatt et al. [49] to model wartime negotiations. Finally Veltman et al. [87] test zeroth-order (no ToM, a game-theory model not taking opponent's actions into account), first-order and second-order ToM on a specific game, called the mod game. They note that in testing, a zeroth-order ToM does not produce different decisions from a first-order ToM (the order of ToM considered throughout this thesis), for this specific game.

The literature review reveals that game theory-based models do not generalize well, and can only be applied to simplistic game-like scenarios. This is also observed by Langley et al. [53]. However, it is arguable whether game-like scenarios are a good measure for ToM. Humans do not necessarily engage in ToM when playing games, as it is cognitively demanding to simultaneously reason about the opponent's mental state, and the effect that opponent's actions might have on the game state [39]. Thus, it stands to reason that employing ToM is something that is not strictly necessary to do well in playing certain games. The result found by Veltman et al. [87] (No difference in decisions made by models without ToM and models with first-order ToM for the game researched in their paper) further validates this hypothesis. Therefore, performance in games does not necessarily

correlate with having ToM, and is therefore not necessarily a suitable measurement of ToM. This, combined with the fact that most game theory models do not generalize well, puts into question whether game theory is a solid foundation for developing aToM.

2.2.3 Cognitive architectures

The term 'cognitive architectures' is an umbrella term for any architecture that attempts to model human cognitive processes [18]. Therefore this term can be used for many of the papers discussed in other categories. For instance, Winfield [96] uses a hybrid cognitive/inverse reinforcement learning architecture that is partially based on cognitive theories, but also simulates the agent, the environment and other actors using machine learning-based strategies. This section will therefore be limited to papers explicitly attempting to model human cognitive processes in a rule-based manner.

In the literature review we find four different papers (Bosse et al. [15], 2 papers from Sarkadi et al. [81, 82], and Panisson et al. [67]) incorporating a rule-based architecture based on the belief-desire-intention (BDI) framework [16]. This BDI framework is widely used in cognitive sciences and is a useful framework to model goal-directed behaviors. The idea behind the framework is that an agent (human or artificial) has a personal set of beliefs; their perception of the world. They also have a set of desires; what they want to achieve. These desires can be based on external stimuli, i.e. their beliefs, as well as internal stimuli. In order to fulfill their desires, the agent will form a concrete set of plans; their intention. The four implemented BDI models are highly abstract, requiring a researcher to define beliefs, desires, intentions and their relations for a specific problem. Therefore the models presented in these four papers [15, 81, 67, 82] do not generalize at all. A fifth paper, by Vossen et al. [89], implements a similar BDI architecture, but in the context of conversational agents. Their model attempts to store information related to a conversation as either beliefs, desires and/or intentions.

Although the BDI framework is a useful framework for modeling goal-driven behavior [16], it is overtly simplistic for creating a model possessing generalizable aToM [53]. Although rule-based architectures, like those based on the BDI framework, can contribute to aToM research, the argument can be made that human ToM is not well-enough understood currently. Therefore we can not replicate human ToM artificially using purely rule-based architectures. However, rule-based architectures are not the only architectures used for aToM research. The five subsections remaining in this chapter will look at machine learning-based architectures.

2.2.4 Observational reinforcement learning

Besides rule-based architectures, several machine learning algorithms are used in aToM research [33]. Reinforcement learning (RL) [93] is a machine learning method that teaches an agent a desired behavior using rewards and/or punishment. This agent operates in a simulated environment [48], for instance a 2D gridworld. We can view the model as separate from the agent [33], and therefore we can view the model as having a ToM, observing and reasoning about another entity (the agent). For

instance, the model can have a different set of beliefs from the agent [93]. The agent can for instance only be aware of a subset of the simulated environment, whereas the model has knowledge of the whole environment. Reinforcement learning models learn using a reward function, and therefore suffer from the same problems as game theory based models, being highly dependent on the choice of reward function. However, several RL techniques exist that bypass the reward function entirely (also see inverse reinforcement learning below). Imitation learning is a RL technique, wherein a model learns to reproduce behaviors of agents acting in a simulated environment. The model will simply reproduce behaviors based on how similar scenarios are to scenarios it has seen before. This bypasses the need for a reward function. However, this makes imitation learning models highly depended on already having seen a certain situation, and makes these types of models unable to generalize well to novel scenarios [33].

In the literature review we find several papers using some form of RL. Some of these will be discussed below (see subsections 2.2.5 and 2.2.6). However, two papers were found that implemented some form of imitation learning. Oguntola et al. [62] created a deep reinforcement learning model that is based on behavioral cloning [86], a subcategory of imitation learning. They create a three-part model, modeling beliefs, desires and actions, according to the BDI framework [16] (for more information on the BDI framework, see subsection 2.2.3). They test their model using a simulated search and rescue task, and train their model using 75 trajectories collected from human participants. Nguyen and Gonzalez [60] propose a hybrid cognitive/observational aToM model, based on existing Instance-based learning theory (IBLT). Using IBLT they construct a model that can learn from the observation of other agents. They believe that their hybrid approach mimicks human ToM like rule-based architectures, while keeping the powerful generalization capabilities of reinforcement learning-based architectures. They base their tasks on Rabinowitz et al.'s paper [77], which is discussed in the section below.

2.2.5 Inverse reinforcement learning

Another closely related RL technique, that also doesn't depend on a reward function, is called inverse reinforcement learning (IRL) [59]. In IRL, a model attempts to estimate a reward function, based on observed behaviors. This reward function can then be used to train a regular reinforcement learning model, allowing it to reproduce the observed behaviors. This technique differs from the imitation learning discussed above, in that it combines observation with inferences about an agent's intentions and beliefs [33]. Therefore it is more suited to infer (false) beliefs and intentions, than the imitation learning mentioned above. Both imitation learning and inverse reinforcement learning are highly suited to solving different types of ToM tasks. However, they require relatively large amounts of training data and computing time. Furthermore, their results are difficult to interpret [53]. Although they are more suited towards a wider set of ToM tasks than game theory-based or cognitive models, it can still be hypothesized that tasks requiring the mechanism of 'shared world knowledge' are the most difficult for reinforcement learning-based models, due to their reliance on a singular reward function.

During the literature review two models were found implementing IRL-based architectures. Winfield [96] designed a hybrid cognitive/simulation-based model for aToM. The architecture equips an agent

with an internal model of itself, the environment and other dynamic actors. For these actors the robot can simulate their behavior in order to anticipate it. At the heart of the architecture are the action and consequence evaluators. These attempt to find which actions lead to desirable outcomes using mechanisms similar to IRL. Winfield performs several experiments with robots, focusing on how robots can safely interact with humans. For instance, one experiment focuses on how a robotic agent can safely navigate moving through a crowd. The second found is by Rabinowitz et al. [77]. Their model makes use of reinforcement learning using a deep neural network. Although their implementation is different to IRL, they are able to learn in a similar fashion to IRL-based architectures: By estimating a reward function, based on observed behaviors of an agent. They define two specific types of ToM: agent-specific ToM is what their model learns in order to model future states of specific agents, and general ToM; the 'meta-learning' that their model does to become better at modeling future states of agents in general. Rabinowitz et al. test their model using a modified version of the Sally-Anne task (see subsection 2.1.3) in a 2D gridworld environment. They have several boxes in this gridworld, and construct several agents to act as different versions of a 'Sally'. For instance, they have agents moving randomly through the gridworld, or agents that prefer a specific box. The model is expected to reproduce this behavior. Rabinowitz et al. reports a high accuracy for their model.

2.2.6 Bayesian inference

In order to lessen the reliance on a singular reward function, Bayesian models instead opt to reason about intention, and model uncertainty by considering the likelihood that an agent holds some belief and has some intention. In order to do so, they consider multiple combinations of beliefs, desires and intentions (in line with the BDI framework, see subsection 2.2.3) and calculate the likelihood of each pairing using Bayesian inference [33]. Usually Bayesian inference is used in combination with some form of reinforcement learning. They are able to reason about a wide variety of tasks. This makes it highly probable that these types of models can reason well about all three mechanisms defined in subsection 2.1.4. However, these models are computationally expensive, meaning they are currently only able to reason about small-scale problems, that reduce the number of possible intentions and beliefs to a minimum [53].

Both review papers [33, 53], as well as some of the papers from the literature review [60, 62] reference research by Baker et al. [10, 9]. For completeness, we discuss this research in addition to the literature review. Baker et al. [10] uses Bayesian inference to infer beliefs, desires and the possible world state. They test their model in an experiment in a 2D gridworld containing two out of three possible food trucks. One of the two food trucks is obstructed from the view of the agent, while the other is not. Using this setup the model can make inferences as to what food truck the agent prefers, and what food trucks the agent believes are in the gridworld.

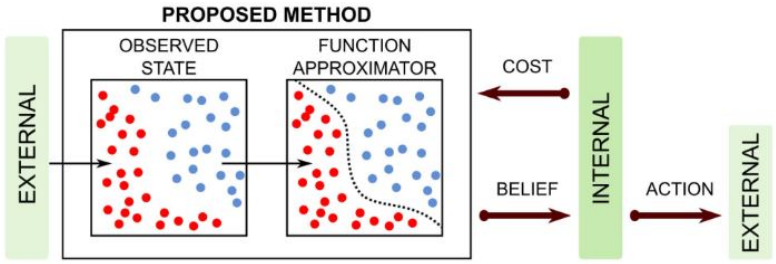

Figure 2.2: Schematic depiction of the proposed model by Patacchiola et al. [69], based on intrinsically motived reinforcement learning (IMRL). The architecture contains an internal layer that uses Bayesian inference to approximate cost (negative reward) functions, which are fed into a RL algorithm that uses these cost functions to generate possible beliefs. These beliefs are then in turn translated into a set of actions by the internal layer. Figure taken from [69].

The literature review also reveals one paper implementing some form of Bayesian inference. Patacchiola et al. [69] developed a model based on intrinsically motivated reinforcement learning (IMRL). This architecture combines Bayesian inference and regular RL in order to model intrinsic motivation, which is difficult to model using only RL. For a more detailed explanation of the model, see Figure 2.2. They test their model on several tasks. In one task, the model has to locate a sticker, which can be in one of several locations. Two informants tell the model where the sticker could be. However, only one of these informants can be trusted. Through repeated experiments the model learns which informant can, and can't be trusted. The second experiment also involves reliable and unreliable informants, but now the model has to learn the name of several objects. In this experiment the model has to concurrently learn the names of objects, as well as learn which informant is reliable, and which informant is not.

2.2.7 Literature review summary

In this section, we evaluated 15 papers implementing an aToM model, and summarized the findings of two review papers [33, 53]. The focus of this review was to evaluate what architectures are used in aToM research, as well as review their strengths and weaknesses. We will now summarize the findings, and see if we can find similarities between papers. We found five types of architectures. Two types of rule-based architectures; game theory-based architectures and cognitive architectures, and three types of machine learning architectures; observational reinforcement learning, inverse reinforcement learning and Bayesian inference. Although the rule-based architectures make up the majority of the papers reviewed (10 out of 15 papers), they do not generalize, and are highly specific to the problem they are designed to solve. Because the models in these papers do not generalize, but rather are designed with specific problems in mind, they can not be evaluated outside of the specific problems they are meant to solve. Because these models are inherently connected to the problems they are designed to solve, they are unsuitable for the research performed in this thesis.

All of the machine learning-based architectures do generalize well, requiring much less specific for-

matting of the problem tasks. All of the reviewed machine learning papers make use of some form of reinforcement learning. Specifically, they all operate using Markov Decision Processes (MDPs) [93]. MDPs are, in mathematical sense, a 3-tuple $< S, A, T >$. MDPs consist of states S, and have an agent that exists in a certain state s. For each state s, the agent will have a set of actions it can perform A_s. This set of actions is often the same for each state in RL, but it is not a requirement. Each action has an associated probability to move the agent to a new state s'. We call this probability the transition function $T(S, A, S')$. As each action is taken at a certain timestep, we can create a global clock $t = 1, 2, ...$ and define $T(s_t, a, s_{t+1})$. It is important to note that all transition functions for a certain state s and action a should sum to 1. To illustrate, consider the example of modeling a chessboard. Each square on the board is considered a separate state s. We could then, for instance, model 4 possible actions a for each state: moving up, down, left and right. Of course, on the edges of this board, some of the actions would not be possible. In this sense, we can model real-world scenarios as MDPs.

Another similarity is that many of the papers extensively make use of the BDI framework (see subsection 2.2.3). The BDI framework [16] is a commonly used method to formalize goal-driven behavior. Many of the tasks evaluated in subsection 2.1.3 show scenarios in which the actor displays goal-driven behavior. Therefore the BDI framework is a useful tool in modeling ToM-related tasks. It has also been shown that the BDI can easily be linked to the structure of RL architectures [43]. We will use the found similarities (MDPs and the BDI framework) in order to discuss requirements for aToM models, and to design methods to evaluate aToM models on the mechanisms found in subsection 2.1.4.

2.2.8 On the emergence of aToM in Large Language Models

During the writing of this thesis, several papers have been published [50, 55] reporting the emergence of ToM-like capabilities of Large Language Models, such as GPT-4 [64]. Large Language Models would offer many advantages in terms of the format of task they can reason about. Many human ToM task are presented as a story (see subsection 2.1.3), or in a visual format, which is also accepted by the latest version of GPT-4 [64]. However, research [50, 55] concerning the emergence of ToM in large language models is not yet peer-reviewed, and concerns proprietary software. Furthermore, no papers were found during the literature review that created a large language model for aToM research. Therefore this thesis will not take large language models as a possible ToM architecture into account.

Chapter 3 - Method

3.1 Formalizing ToM mechanisms

Now that an overview of human ToM tasks and mechanisms is present (see subsections 2.1.3 and 2.1.4), as well as an overview of the most common architectures and tasks used in aToM research (see section 2.2), we can delve into what these mechanisms entail, and how we can evaluate aToM model's performance on these mechanisms. We will start by formulating how humans employ the three mechanisms found in subsection 2.1.4 (interpreting actions, visual perspective-taking and shared world knowledge) in order to solve tasks related to these mechanisms. We will also delve into what auxiliary skills are required in order to solve tasks related to these mechanisms. ToM as a skill can not be measured directly [94], thus additional skills, not necessarily related to ToM, are required for a participant to successfully reason about these ToM tasks. In order to distinguish between ToM skills and auxiliary skills we make use of the criteria defined in subsection 2.1.2. If a skill does not fullfill either the nonmerging or mentalizing criteria, that skill can be considered to be an auxiliary skill.

The idea of a formalized rule-based approach for human learning and reasoning has been thoroughly explored in literature [35], and can be linked to the study of Theory of Mind through the framework of theory theories [5]. Theory theories state that humans learn about the world by actively building theories about concepts, which they then can later use to reason about said concepts. A formalized rule-based approach can both be used to further understanding of human learning through computer analysis [34] as well as be used to develop AI that thinks and reasons in a humanlike manner [52].

3.1.1 Assumptions and definitions

In subsection 2.1.4, we categorized the tasks found by Quesque and Rossetti [76] and Byom and Mutlu [17] into three mechanisms. The following subsections will create three formal frameworks, one for each mechanism. In order to so, we will dissect an example task for each mechanism, giving an in-depth explanation for how a human participant would solve said task. Thus, we operate on the assumption that all tasks categorized in subsection 2.1.4 under a mechanism, are solved in a similar manner. However, one can verify each of the frameworks accuracy, by applying it to any of the other tasks categorized in a mechanism. After each of the three frameworks have been established, we can reason about the three mechanisms' similarities and differences.

We start by creating a vocabulary of concepts that are relevant to ToM. These are listed below. There are two types of concepts. The first category comprises base concepts, for which definitions are provided within the context of ToM tasks. In the second category are ToM concepts. We will base these ToM concepts on the BDI framework, as this framework is both used in the study of

human ToM[1], as well as being commonly used when creating aToM models (see section 2.2).

- **Base concepts**
  - Task: A task that tests the ToM ability of a participant (see below). A task consists of a presented scenario, followed by one or more questions to the participant. The task can be in visual, audio, and/or written format.
  - Participant: The person whose ToM skills are being evaluated in the task.
  - Actor: Any person present in the task's scenario. The participant's ToM skills are evaluated by reasoning about actors.
  - Object: All physical items that are present/mentioned in a task, but are not an actor.

- **ToM concepts**
  - Belief: Actors will hold beliefs that are personal to them, and that might be different from reality. Participants should be able to understand and reason about what actors believe from the scenario presented in the task. In case of the participant, we distinguish two types of beliefs:
    - Task-based beliefs: Beliefs that were inferred from the task's scenario.
    - Knowledge: All held beliefs by the participant before partaking in the task. It is important to note that these beliefs might still be different from reality, and can differ from participant to participant.
  - Desire: The motivation of an actor. Desire is different from intention, in that desire is the overall motivation of an actor, whereas intention is more specific and goal-oriented, outlining the specific actions one wants to take.
  - Intention: What an actor intends to do. Intention is always related to a specific action, unlike desire. Pylyshyn [73] refers to this intention as: having a relationship to a state of affairs and gives examples: wanting something, needing something, thinking about something (see section 2.1.1).
  - Action: Actions, which in this context also include statements and expressions, are performed by the actors and can be observed by the participant. They are the only of these ToM concepts that can be observed by the participant.

If we take the case of a single human taking an action in isolation (without any relation to ToM), we can assume these concepts relate to each other in the following manner: An actor has a desire, and a set of personal beliefs. This desire and beliefs lead to an intention, which results into an action (see Figure 3.1). For example: I'm hungry and believe there is food in the fridge. I want food, because I'm hungry. Therefore I will get food from the fridge. In this examples, we have a desire

[1]The BDI framework is related to theory theories (see subsection 3.1), which is one of the two dominant approaches in explaining how humans reason about ToM-related concepts (see subsection 3.1). The other dominant approach is called simulation theories, which states that humans run simulations of themselves from the perspective of another person, in order to infer their state of mind.

(being hungry), and have a belief (there is food in the fridge). This desire and belief leads us to our intention (wanting to get food from the fridge), which leads to the action of getting food from the fridge.

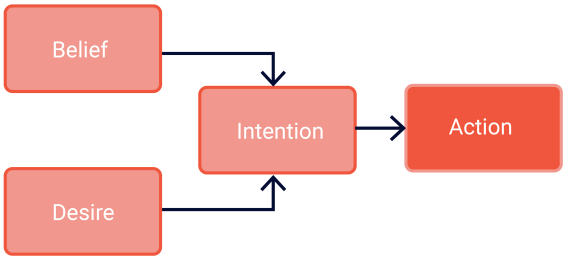

Figure 3.1: Decision making model for a single person using the concepts of desires, beliefs, intentions and actions. A person has a desire, and a set of beliefs. These beliefs and desires inform their intention, which consecutively results in an action.

3.1.2 Interpreting actions

We will start by formalizing how a participant reasons about tasks relating to the mechanism of interpreting actions. All the tasks categorized in subsection 2.1.4 under 'interpreting actions', are false belief tasks. Therefore, we take a look at an example false belief task, namely the classic Sally-Anne experiment [12]:

- **Interpreting actions/False belief task [12]**
  - Scenario: Sally puts a marble in a white box, then leaves the room. Anne takes the marble from white box and puts it in a black box, then Sally comes back and looks for her marble.
  - Question: Where will Sally look for her marble?
  - Answer: She will look in the white box, because she believes her marble to be there.

We will now use the established ToM concepts (see subsection 3.1.1) to break down this task. To start with, the scenario provides us with an **action:** 'Sally looks for her marble'. As stated in concept's definition, an action is the only concept that can be observed by the participant. Furthermore, Sally's **desire** and/or **intention** can be inferred from the question: Sally wants her marble (desire), and therefore she will look for it (intention). If we look at the decision-making model for a single actor (see Figure 3.1), we can see that the actor's **desire**, **intention** and a possible **action** are already covered by the task scenario. This leaves us with the **beliefs** of the actor. In the example task, the participant is presented with the following **task-related beliefs:**

1. Sally put the marble in the white box.

2. While Sally was not present, Anne took the marble from the white box and put it in the black box.

3. Therefore, the marble is now in the black box.

It is important to note, that these presented **task-related beliefs**, are those of the participant, not of the actor (Sally). Therefore, in order to successfully solve this task, the participant has to be able to separate their own **task-related beliefs**, from those of the actor. The participant has to reason that, because Sally was not present while the participant learned belief 2 and 3 mentioned above, her personal **beliefs** will only consist of belief 1. Therefore the framework for the mechanism of interpreting actions will look as follows: The **participant** uses their **Theory of Mind** to reason about a given **actor's action**, by separating the **actor's personal beliefs** from their own **task-related beliefs**. See Figure 3.2 for the framework. Figure 3.3 shows how the example task given in this section relates to the framework. It is worth observing that this separation of beliefs corresponds to the nonmerging criteron found by Quesque and Rossetti (see subsection 2.1.2).

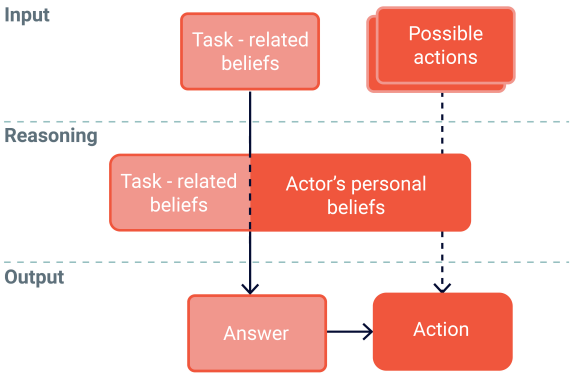

Figure 3.2: Framework for solving tasks related to 'interpreting actions'. Beliefs and actions relating to the actor are in dark red. Beliefs related to the participant are in light red. The participant is given an actor's (possible) action, and has a set of task-related beliefs. The participant uses ToM to separate the actors beliefs from his own task-related beliefs. He then reasons about the given action and separated belief to answer the question.

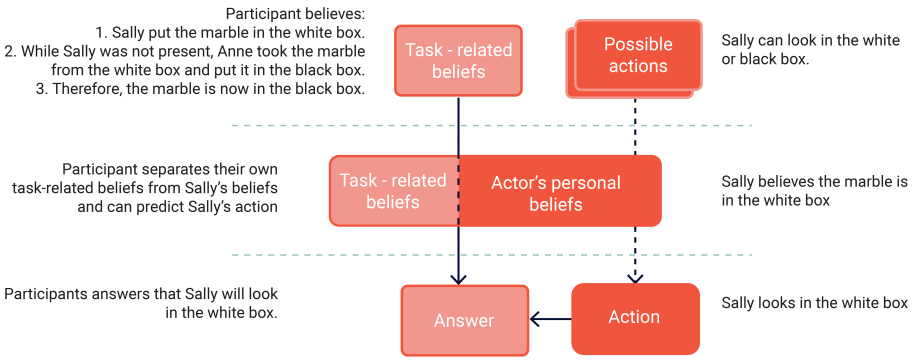

Figure 3.3: This figure shows how the example task from section 3.1.2 relates to the framework presented in Figure 3.2.

Auxiliary skills In order to solve the example task, the participant needs to be able to understand the scenario, as well as the question, which can either be presented in written form or be told by a researcher. Therefore, for this specific scenario, the participant needs some language comprehension skills, as well as the ability to mentalize the concepts presented in the scenario (as per the mentalizing criterion - see subsection 2.1.2).

This Sally-Anne task has been traditionally used in the context of developmental psychology. In experiments it was noted that children below the age of four are on average not able to successfully reason about this task [12]. However, additional research on the on the relation between language acquisition and development of ToM skills puts this result into perspective. Hale and Tager-Flusberg [36] experimented on sixty preschoolers. All preschoolers first performed a false belief test and failed it. The preschoolers were then divided into three groups. Each group then had training in one of three categories. One group received false believe task training. A second group received training on sentential complements[2] and a third group received training on relative clauses[3] as a control. After the training they were given a different false belief task. The group that was given false belief training now performed better on the test, as was to be expected. The control group that was trained on relative clauses, did not perform better of the false belief task. However, the group that was trained on sentential complements performed similar to the false belief training group. This finding suggests that, at least for false belief tests, some form of language-related skills can be of great use in enhancing ToM.

[2]Sentential complements are subordinate clauses that function as objects, subjects or complements to the main clause, e.g. 'She believes **that he is innocent.**' or '**If it rains,** we'll stay indoors.'

[3]Relative clauses are dependent clauses that provide description to the noun they modify, e.g. 'The car, **which runs on electricity.**' or 'The woman **who lives next door.**

3.1.3 Visual perspective-taking

We will now proceed with another mechanism found in subsection 2.1.4. Visual perspective-taking has three different tasks associated with it: The spatial orientation task [40], the perspective-taking task [71] and the director's task [98] (see subsection 2.1.3). We will use the spatial orientation task as a basis for the visual perspective-taking framework. Please refer to Figure 2.1 for the task example.

We need to make some slight modifications to the task to allow framing using the concepts provided in subsection 3.1.1. The original task asks the **participant** to imagine themselves in the scenario presented in Figure 2.1. During the original explanation of the task (see subsection 2.1.3), we've already established that we can view the simulated self as a separated entity from the participant. Therefore we will refer to the simulated self as the **actor**.

Again, in this task, we are presented with an **action**: 'Point at the traffic light'. In this task, **desires** and **intentions** are relevant to the same degree, as in the Sally-Anne task studied in the previous subsection (see subsection 3.1.2). The participant has to point at a location, for which it is given that the actor (the simulated self) will do so as well. Why this actor wants to do so, is irrelevant to solving this task.

This leaves us with the concept of **beliefs**. Although some **knowledge** is required, namely which object in the figure represents a stop sign, a house and a traffic light respectively, this is not relevant to the part of the task that constitutes ToM. We can prove this relevancy, by verifying that 'recognizing a stop sign from a set of pictograms' as a task, does not fulfill the nonmerging, nor mentalizing criteria established in subsection 2.1.2. Therefore, **knowledge** does not play a role in reasoning about the ToM-related portion of this mechanism, but rather is an auxiliary skill. We are then left with the concept of **task-related beliefs**. In order to solve the spatial orientation task, the **participant** has to synthesize the world in which the actor exists, in order to learn what the **actor's personal beliefs** are. Similar to the interpreting actions framework, these **actor's personal beliefs** are separate from the **participant's task-related beliefs**, in that the perspective of the participant's beliefs differ from those of the actor.

Therefore we can conclude that the framework for reasoning about the mechanism of visual perspective-taking is the same as the framework for interpreting actions (see Figure 3.2). The participant is provided with a (possible) **action** (pointing in a direction), and asked to specify it (by pointing in the correct direction). In order solve this task the participant is provided with a set of **task-related beliefs**, and has to separate the **actor's personal beliefs** from their own.

This leaves us to wonder whether the distinction between 'interpreting actions' and 'visual perspective-taking' as mechanisms exists. Indeed, it seems that the main difference between 'interpreting actions' and 'visual perspective-taking' is in the auxiliary skills required for the task. When we look at a different task: the director's task (see Figure 3.4), this similarity becomes even more apparent.

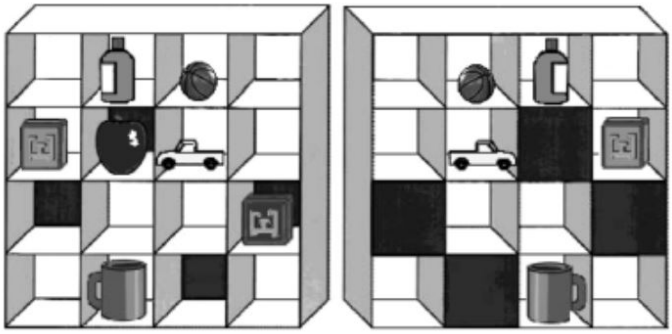

Figure 3.4: Example scenario for the director's task. The participant views the cabinet from the perspective shown in the left image. An actor stands behind the cabinet, so that they see the cabinet from the perspective shown in the right image. The participant is then asked by the actor to move the block with the letter E on it, up one square.

As stated in Figure 3.4, the participant is asked by an actor, that sees the cabinet from a different perspective, to move the block with the letter E on it, up one square. In order to solve this task, the participant has to separate the **actor's personal beliefs** (there is one block with an E on it), from their own **task-related beliefs** (there exist two blocks with an E on it, one of which is hidden from the backside). No reasoning about **desires** or **intentions** is required to solve this task. Therefore, the mechanisms of 'interpreting actions' and 'visual perspective-taking' are identical, at least in terms of ToM reasoning.

Auxiliary skills However, these mechanisms do differ in the auxiliary skills required for solving tasks related to them. Tasks related to 'interpreting actions' are text-based, being either written or told by a researcher. Tasks related to 'visual perspective-taking' are all visual, and require the participant to have a certain level of spatial awareness or spatial orientation skills. Simulation theories is, together with theory theories (see subsection 3.1), one of the leading theories for how humans reason about ToM. Simulation theories state that humans simulate themselves from the perspective of another person in order to infer the mental state of that other person. Simulation theories might be especially applicable to the mechanism of visual perspective-taking.

3.1.4 Shared world knowledge

This leaves us with the final mechanism, that of shared world knowledge. For shared world knowledge, we can look at an example scenario from the 'strange stories' task [38] (see subsection 2.1.3).

- **Shared world knowledge/Strange stories task [38]**

- Scenario: Katie and Emma are playing in the house. Emma picks up a banana from the fruit bowl and holds it up to her ear. She says to Katie, "Look! This banana is a telephone!"
- Question: Why does Emma say this?
- Answer: Because Emma is playing with Katie and pretending the banana is a telephone, even though she knows it is not a telephone.

Again, in this task, we are given an **action**, performed by an **actor**: Emma picks up a banana and holds it to her ear. She then says that the banana is a telephone. Unlike in the tasks related to 'interpreting actions' and 'visual perspective-taking', **desires** and **intentions** are equally as important as **beliefs** in order to answer the question. We can come up with different combinations of beliefs, desires and intentions that would lead to the described action, but to a different answer. For instance, a wrong answer would be: Emma believes that the banana is a telephone, she wants to call her aunt (desire), and thus she picks up the banana to call her aunt (intention). This is a logical answer in the sense that in would explain the action described in the scenario, and yet it is the incorrect answer.

It is also important to note that, unlike in the false belief task, the only **task-related belief** the participant has, is that Katie and Emma are playing. The rest of the scenario solely describes the **actions** Emma took. Therefore the interpreting actions/visual perspective-taking framework from Figure 3.2 can not be applied to the strange stories task.

Instead, the participant has to rely on their **knowledge** to answer the question. The amount of knowledge a participant has, is not something that can feasibly be itemized, even for the participant themselves. Instead, the participant has to invert the problem. This can be done by weighing multiple possible answers on how well they explain the given **action**, based on the participant's **knowledge**. This inverted manner of problem solving in human cognition is well documented, and can be modeled using Bayesian inference [85].

For instance, so far there are two possible explanations given for Emma's actions:

1. Emma **believes** that the banana is a telephone, she wants to call her aunt (**desire**), and thus she picks up the banana to call her aunt (**intention**).
2. Emma is playing (**desire**) with Katie and pretending (**intention**) the banana is a telephone (by picking it up and holding it to her ear), even though she knows (**belief**) it is not a telephone.

Even though explanation 1 is a valid answer, the participant can use their **knowledge** to deduce that it is unlikely that Emma believes that the banana is a telephone, as it is highly likely that she knows the difference between the two objects. Explanation 2 is much more in agreement with the **knowledge** of an average participant, as it is a well-known fact that during play, children engage in pretend-play, during which they create a false reality for the sake of enjoyment.

Therefore we arrive at the following framework for the mechanism of shared world knowledge: The **participant** uses their **Theory of Mind** to generate multiple possible explanations in the form of

beliefs, desires and intentions for a given **actor's action**, and then chooses the most likely set of **beliefs, desires and intentions** by weighing the likelihood of each set against their combined **task-based beliefs** and **knowledge**. See Figure 3.5 for the framework. Figure 3.6 shows how the example task given in this section relates to the framework.

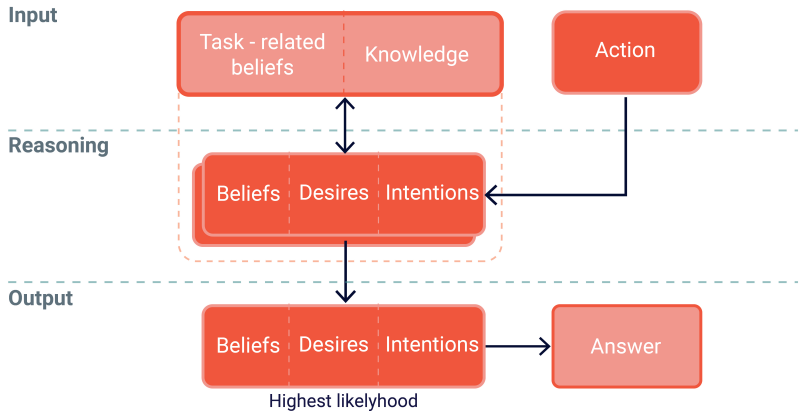

Figure 3.5: Framework for solving tasks related to 'shared world knowledge'. Beliefs, desires, intentions and actions relating to the actor are in dark red. Beliefs related to the participant are in light red. The participant is given an actor's action, and has a set of task-related beliefs, as well as prior knowledge. The participant uses ToM to generate a set of possible beliefs, desires and intentions the actor might have, and compares the likelihood of each of these against his own beliefs. The most likely explanation is then given as an answer.

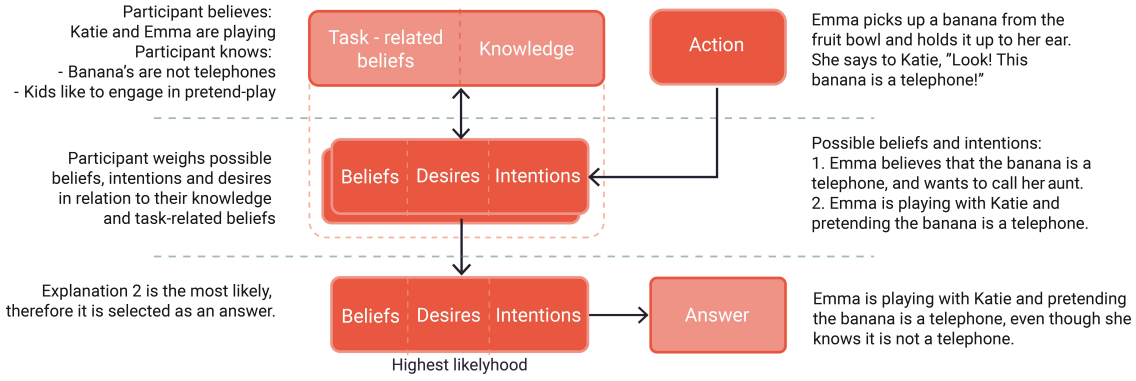

Figure 3.6: This figure shows how the example task from section 3.1.4 relates to the framework presented in Figure 3.5.

Auxiliary skills Like the false belief tasks examined in section 3.1.2, most of the tasks categorized under the mechanism of shared world knowledge are either written, or being told to the participant by a researcher. Therefore the same auxiliary skills apply: The participant needs some form of language comprehension skills in order to solve tasks related to shared world knowledge. It should be noted, that for this mechanism, some tasks were found [23, 80] that presented the task in a visual format (either through a video, or through a comic strip). However, unlike for the mechanism of visual perspective-taking, no spatial orientation skills are required, as the participant is mostly asked to reason about the concepts provided in the framework (see Figure 3.5), and is not asked to reason about any spatial concepts.

3.2 Reconceptualizing ToM mechanisms for artificial intelligence

The previously defined frameworks, shown in Figures 3.2 and 3.5, describe how humans solve ToM tasks for interpreting actions, visual perspective-taking and shared world knowledge. These frameworks are a useful tool in order to distill what type of reasoning is used in the ToM tasks found in subsection 2.1.3, allowing us to design new tasks that test the same mechanisms for human participants. However, being able to design new ToM tasks for humans does not necessarily equate to being able to design tasks to evaluate artificial ToM models. This is, because of the auxiliary skills required for each mechanism. Human ToM tasks are presented in either a written/verbal format (in case of tasks relating to interpreting actions/shared world knowledge), or in a visual format (in case of tasks relating to visual perspective-taking). Artificial models are unable to reason about written/verbal or visual tasks. Therefore, we need to reconceptualize how we present ToM tasks for artificial models. In section 2.2, we discussed several of the most commonly found architectures used in aToM research. These architectures were either rule-based (game theory-based architectures, or cognitive architectures), or had some machine learning component (observational reinforcement learning, inverse reinforcement learning or Bayesian inference). This section will discuss how we can reconceptualize the frameworks found in the previous section in order to design tasks for aToM models. We will also discuss the limitations of the found architectures, in order to find what is required of aToM models in order to succesfully reason about all three types of ToM mechanisms.

3.2.1 On rule-based architectures

As discussed in subsection 2.2.7, both types of rule-based architectures found during the literature review (game theory-based architectures and cognitive architectures) are build with specific tasks in mind, and do not generalize to other types of tasks. The goal of this thesis is to formally define the different mechanisms relating to ToM, as well as to evaluate aToM models' effectiveness in reasoning about these mechanisms. Because these rule-based architectures are only designed for a single task, these type of models cannot be evaluated on other types of tasks. Therefore the subsections below will mainly focus on the machine learning-based architectures, as these types of architectures are able to generalize to other problems with minimal modification (an example of which can be found

in chapter 4). We will however discuss how one can design new rule-based aToM models making use of the frameworks discussed in section 3.1. Especially, new cognitive architectures based on the BDI framework [16] could be designed with relative ease, as the frameworks from section 3.1 all make use of concepts from the BDI framework.

3.2.2 Visual perspective-taking

We will start with the mechanism of visual perspective-taking, as humans are tested on this mechanism using spatial tasks (see subsection 3.1.3). During the aToM architecture literature review, it was found that the most common architectures that generalize well (machine learning-based architectures) all make use of MDPs (see subsection 2.2.7). A common manner to visualize MDPs is a 2-dimensional gridworld [93], a coordinate system along two dimensions. Through the use of these gridworlds, these type of models can reason very well about spatial tasks [93].

This leaves us with the question how we can adapt the framework from Figure 3.2, in order to allow us to design a ToM task for artificial models. Tasks relating to visual perspective-taking provide the participant with a set of **task-related beliefs** and a set of **possible actions**. They are then asked to provide the correct **action** from the set of **possible actions**. Thus, how can we model **beliefs** and **actions** using the gridworld environment?

In his paper, Jara-Ettinger [43] proposes how to reconceptualize the BDI framework for inverse reinforcement learning. He proposes the following reconceptualizations (see Figure 3.7):

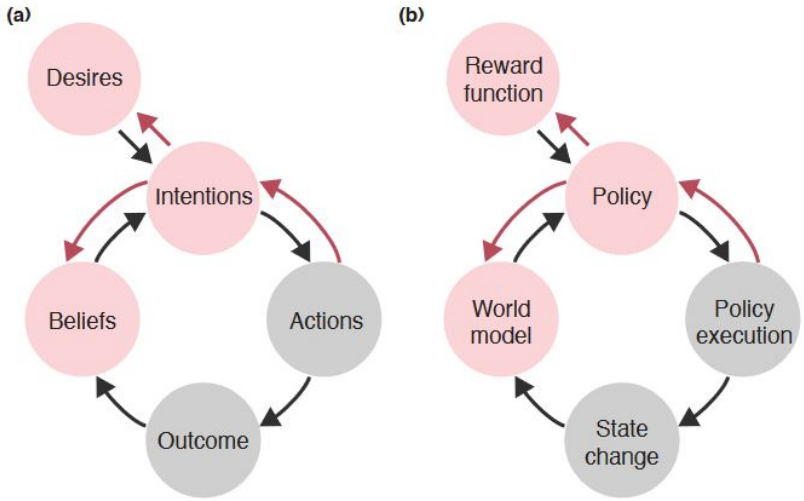

Figure 3.7: Reconceptualization of the BDI framework as inverse reinforcement learning. Image taken from [43].

- **Beliefs:** Can be reconceptualized as the environment, i.e. the statespace in the MDP/gridworld.

- **Desires:** Can be reconceptualized as the reward function, informing the decisions made by the agent in the model (see subsection 2.2.5).
- **Intentions:** Can be reconceptualized as the policy. The policy in (inverse) RL is the mapping of what action should be performed in what state.
- **Actions & outcome:** Do not need a reconceptualization, RL architectures already explicitly model actions.

We will base our reconceptualization on Jara-Ettinger's reconceptualization. For the framework relating to visual perspective-taking (see Figure 3.2), only **(task-based) beliefs**, **desires** and **actions** are relevant. We will explain how we reconceptualize these concepts below.

Actions are the most simple to model, as the structure of gridworlds already includes a parameter to model them (see subsection 2.2.7). Actions in a gridworld define how we can go from one state to another. Thus, taking the example of the spatial orientation task [40] (see subsection 2.1.3), we could, for instance, model the possible actions as {pointing up, pointing down, pointing left, pointing right}, with each action leading to a corresponding state. The model then has to choose the correct action from this list of actions, in order to solve the task.

How can we then model **task-related beliefs** in a gridworld environment? For this we can leverage how we represent the states. We could, for instance, model some states to include an object (like the stop sign, house or traffic light from the example task shown in subsection 3.1.3). Therefore we can represent the **task-related beliefs** as the specific configuration of the environment.

Thusfar, we have established how to model all necessary inputs, in order for an aToM model using gridworlds to solve tasks related to visual perspective-taking. We can provide various configurations of **beliefs**, i.e. the location of the objects in the environment, for which a corresponding correct **action** exists. However, one problem remains: How does the model know which action corresponds to which configuration of beliefs? In other words, how does the model know where the agent should point? As stated in section 3.1.3, the formulation of the question represents the concept of **desire**, at least for tasks relating to visual perspective-taking. Jara-Ettinger [43] models this **desire** as the reward function, used in IRL. However, during the literature review (see section 2.2), several types of RL-based architectures were found, and not all of these architectures make use of a reward function. Therefore we propose to model the concept of **desire** in a more broad sense.

In order to understand how we can model **desires**, we first need to understand how machine learning models learn. Machine learning models are trained on a labeled dataset. For the architectures researched during the literature review (see section 2.2), this means that they are presented with a dataset containing examples of agents acting in various configurations of the environment. If the model is able to make the correct inferences, they are then able to repeat the correct behavior, even if they have never seen a specific configuration of the environment before. Thus, we could teach a model what is needed to answer the question in a variety of task-related beliefs during training. Then during testing, we can present the model with a novel configuration of the environment, and it should provide us with the correct action. Thus, we can model **desires** as the data provided during training. IRL architectures infer a possible reward function during training, therefore for the case of IRL-based architectures, this is approximates modeling **desires** as the reward function.

However, this approach does have one disadvantage: The model has to be retrained if we want to change the question (**desire**). In the example task from subsection 3.1.3, the participant is asked to point at the traffic light, if they are standing at the stop sign and are facing the house. If we, for instance, want to change this question to pointing at the cat (also present in the scenario) instead, the model would have to be retrained on a different set of behaviors. Therefore, most of the found implementations during the literature review (see section 2.2) generalize poorly to the concept of **desire**. In order to better solve these types of tasks, it is recommended to create an architecture in such a manner that it generalizes to be able to answer a variety of questions. This could be accomplished by explicitly modeling a **desire** parameter.

On rule-based implementations Most cognitive architectures based on the BDI framework actually fare better in terms of generalization towards the question asked. This is, because these architectures explicitly model desire as a parameter, allowing a researcher to define different desires (e.g. desiring to point at the stop sign/traffic light/house/etc.). However, none of the models found during the literature review (see section 2.2) model any spatial relations. Therefore a cognitive architecture wanting to solve tasks related to visual perspective-taking, would have to explicitly model spatial properties and/or relations. However, this still poses the question, whether these models can generalize to a variety of tasks remains to be seen. What would happen if we for instance change the question to: Walk a path circling the house, and then touch the traffic light. The reinforcement learning-based architectures discussed above would be able to generalize to this type of question without modification to the architecture. Whether a cognitive architecture can also generalize to this type of behavior remains to be seen.

3.2.3 Interpreting actions

In subsections 3.1.2 and 3.1.3, it was found that in terms of ToM reasoning, no difference exists between the mechanisms of interpreting actions, and visual perspective-taking. However, these types of tasks differ in the auxiliary skills required in order to solve them. This is, because tasks related to 'interpreting actions' are written, or told verbally, whereas tasks related to 'visual perspective-taking' are visual, and require the participant to have spatial orientation skills.

It can be hypothesized that, out of the three mechanisms presented in this paper, cognitive architectures (using the BDI framework) should prove most successful in reasoning about the mechanism of interpreting actions. This is, because all cognitive architectures reviewed during the literature review (see subsection 2.2.3) are relational architectures, explicitly modeling beliefs, desires and intentions. Therefore, a similar architecture can be designed to solve tasks related to the mechanism of interpreting actions. However, these types of architectures still require the researcher to explicitly model beliefs, desires, intentions and actions for a specific problem. This leaves the problem of these models not being able to generalize to other problems, even other problems relating to the mechanism of interpreting actions.

If we consider RL-based architectures, the main problem for this mechanism is how we can translate the written task into a task that can be used within a MDP. The previous subsection (3.2.2) already

highlighted how we can reconceptualize the framework from Figure 3.2 for RL-based architectures. This reconceptualization still applies to the mechanism of interpreting actions. Thus, taking the Sally-Anne task as an example (see subsection 3.1.2) we can model the **possible actions** as {Sally opens the white box, Sally opens the black box}. However, how do we model the **task-based beliefs**? In subsection 3.1.2, the following **task-based beliefs** were found for this specific task:

1. Sally put the marble in the white box.
2. While Sally was not present, Anne took the marble from the white box and put it in the black box.
3. Therefore, the marble is now in the black box.

We notice that these three **beliefs** imply some temporal relation: First, Sally put the marble in the white box, then she leaves the room, then Anne moves the marble from the black box to the white box. If we look at other tasks categorized under the mechanism of interpreting actions (see subsection 2.1.4), we find that all of these tasks have some form of temporal relation. Most of the reviewed RL architectures in the literature review (see section 2.2) were found to be unable to model temporal relations, other than that of the agent, i.e. the only object that can move in the environment is the agent. It is therefore recommended that future aToM models include some ability to manipulate the environment in a temporal manner, so that they can model the temporal relations required for solving tasks related to the mechanism of interpreting actions.

However, if we are unable to model the temporal relations required to model the **task-based beliefs**, we can instead model the state of the **task-based beliefs** at the moment in time the question is asked. Thus, for our Sally-Anne example task, this would entail representing two states: One state that represents a box that currently contains a marble, and one state that represents a box that is currently empty, but is marked as the box in which Sally placed the marble. We can then create a training set that includes several examples of the agent (Sally) moving to the box marked as the as the box in which Sally placed the marble (and not to the box that currently contains the marble). In this manner, we can represent the temporal relations necessary to model tasks related to the mechanism of interpreting actions.

To summarize, we can reconceptualize tasks related to the mechanism of interpreting actions for RL-based architectures by directly modeling the concepts and relations given in the written scenario, in a similar manner to how we can model tasks related to the mechanism of visual perspective-taking (see subsection 3.1.3). This manner of reconceptualizing, means that our models can poorly generalize to the question asked (**desire**), needing to be retrained for every change in the specifics of the task. Furthermore, most RL-based architectures found during the literature review (see section 2.2) prove to be unable to modify the environment in a temporal sense. Therefore, the temporal relations presented in the written scenario have to be modified in such a way, that all relevant **task-based beliefs** are included in the environment at the moment the question is asked.

3.2.4 Shared world knowledge

Unlike the mechanisms of visual perspective-taking and interpreting actions, the mechanism of shared world knowledge has a different associated framework (see Figure 3.5). There are three main additions to the shared world knowledge framework (Figure 3.5), as compared to the interpreting actions/visual perspective-taking framework (Figure 3.2):

1. A distinction is made between **knowledge**, and **task-related beliefs**.
2. Both **desires** and **intentions** are included as a concept to reason about.
3. (a) The participant is asked to both generate answers in the form of **beliefs**, **desires** and **intentions**, as well as (b) infer which answer is the most likely according to their own **beliefs** (both **task-based beliefs** and **knowledge**).

In this subsection, we will attempt to address the three additions of the shared world knowledge framework mentioned above, noting what architectures are most suitable for each addition, as well as looking at what is required for a model to reason about the additions mentioned.

Modeling knowledge None of the architectures evaluated in the literature review (see section 2.2) make a distinction between **knowledge**: beliefs held prior to the advent of the task, and **task-based beliefs**: beliefs learned during the task. This distinction is especially irrelevant to rule-based architectures, as these types of architectures do not include a learning component. Therefore all **beliefs**, either **knowledge** or **task-based beliefs** have to be provided by the researcher. However, the machine learning architectures include a learning component during the training phase. In the previous subsections (see subsections 3.2.2 and 3.2.3), we leveraged this training phase to learn the model what behavior it was asked to reproduce, i.e. to model the concept of **desire**. However, we can also use this training phase to impart **knowledge** to the model. For instance, consider the Sally-Anne experiment, discussed in subsection 3.1.2. In subsection 3.2.3 we discussed how we can teach a RL-based architecture how to solve this task. Now consider the case where we swap out either the black or the white box for a glass box. Assume we trained a model on how to solve the Sally-Anne task. Now we provide additional training samples to this model, in order to teach it that glass boxes are transparent. For instance, we could include samples that have one or more glass boxes, one of which contains Sally's marble. In these samples, the agent would open the box containing the marble (as they can see that their marble is inside). This represents the fact that the agent can see the marble if the box is made of glass. In this manner, we can provide the model with additional **knowledge**. If the model is able to correctly learn knowledge, it should then be able to solve Sally-Anne tasks, as well as modified Sally-Anne tasks, where one or either of the boxes is made of glass.

Including desires and intentions For most of the architectures discussed in the literature review (see section 2.2), the inclusion of **desires** and **intentions** as concepts poses little to no added

difficulty in terms of modeling. Cognitive architectures based on the BDI framework explicitly include parameters modeling **desires** and **intentions**. However, these parameters again have to be specified by the researcher. For IRL-based architectures, Jara-Ettinger [43] provided a reconceptualization for **desires** and **intentions** (see subsection 3.2.2). These reconceptualizations can be extended to work for different RL-based architectures. However, on the topic of reconceptualizing **desires**, Jara-Ettinger specifically notes that a singular reward function can't model the different desires an actor might have. This notion is in line with what was found during the literature review (see section 2.2), and was found in previous sections (see subsections 3.2.2 and 3.2.3), with both sections noting poor generalization towards the concept of **desire**. Thus, for RL-based architectures, only a single desire per training set can be modeled.

Inferring the correct hypothesis The last noted addition (addition 3, see the start of this subsection), consists of two parts. (a) The participant both has to generate possible answers (from now on, referred to as **hypotheses**), (b) as well as infer which **hypothesis** is correct. We will start discussing addition 3b: How we can model inferring the correct **hypothesis**. If we look at the models found in the literature review (see section 2.2), we find two specific models that seem suited towards specifically this type of inference: Hiatt et al. [42] (see subsection 2.2.2) created a probabilistic model to consider different hypotheses to explain the behavior of an actor. However, the different hypotheses have to first be known by the model, as it can't generate these hypotheses itself. Furthermore, the authors note poor scalability, stating that the model works best in example scenarios with a limited number of hypotheses. Furthermore, models in the category of Bayesian inference (see subsection 2.2.6) should prove to be suited towards tasks relating to this mechanism. This is, because it has been shown that the manner through which human participants weigh multiple possibilities against their knowledge, can be modeled using Bayesian inference [85]. Specifically, the model created by Baker et al. seems suited towards 'shared world knowledge'-related tasks. This model specifically infers specific parameters related to **beliefs**, **desires** and the world state, using Bayesian inference. However, their model again only considers a bounded number of **desires** and **beliefs**, that have to be explicitly modeled by the researcher, and therefore do not generalize. For instance, in their experiments, they consider a three different food trucks. They consider the possibility that an agent **desires** a specific food truck, and its **belief** of certain food trucks existing at a certain place in the gridworld. Furthermore, they note scalability issues, similar to Hiatt et al. [42]. Although both models note scalability issues, and can't generate **hypotheses** autonomously,we still recommend including some type of Bayesian inference in future aToM models, in order to infer what the correct **hypothesis** is.

Generating hypotheses None of the models discussed in the literature review (see section 2.2) allow for the generation of hypotheses. In his paper, Jara-Ettinger [43] also notes this problem: Humans show excellent cognitive ability to disregard reasoning about **beliefs** that are clearly true (e.g. Sally, from the example discussed in subsection 3.1.2, still remembers where she left the marble once she gets back into the room), or are irrelevant to solving the task at hand (e.g. Sally believes it is raining outside). The same is true for **desires** and **intentions**. However, all of the models discussed in the literature review need a complete representation of all possible **beliefs**, **desires** and **intentions**. This need for a complete representation means that these models can not generate novel hypotheses related to these concepts, but rather have to choose hypotheses as specific combinations

of **beliefs**, **desires** and **intentions** from all possible beliefs, desires and intentions. Thus, current models are not able to generate hypotheses at all, and a researcher has to explicitly specify all possible **beliefs**, **desires** and **intentions**.

Thus, when reconceptualizing tasks related to the mechanism of shared world knowledge for aToM models, currently the task has to be simplified to allow only for a select number of possible **hypotheses**. If we make this simplification, only the two models discussed in the paragraph: 'inferring the correct answer', allow us to reason about tasks related to the mechanism of shared world knowledge. However, Jara-Ettinger [43] provides a recommendation for how we can create RL-based architectures that allow us to better reason about tasks related to shared world knowledge: Thusfar, we've tried to capture all the **beliefs** held by the actor (the agent). However, Jara-Ettinger [43] proposes to create a model that instead holds its own set of **beliefs**, and then only attempts to construct in what ways the actor's **beliefs** differ from their own. Using this difference in **beliefs**, the model can then generate possible **hypotheses**, reasoning about the actor's **beliefs**, **desires** and **intentions**. This recommendation is somewhat similar to the model created by Patacchiola et al. [69], which makes the distinction between an environment representing its own beliefs, and another environment representing the actual (physical) environment. This approach could similarly be extended to separate environments for the differences in beliefs between the participant, and the actor.

3.2.5 Summary

In this section we discussed how we can reconceptualize the frameworks from Figures 3.2 and 3.5. For the mechanisms of visual perspective-taking and interpreting actions, we found how we can represent **beliefs**, **actions** and **desires** for current RL-based architectures. However, we noted that current RL-based architectures generalize poorly to the concept of **desire**. We also provided recommendations for how one can design rule-based architectures based on the BDI framework, for the mechanisms of of visual perspective-taking and interpreting actions. Also, we noted that tasks relating to the mechanism of interpreting actions often include temporal relations in the scenario, and current architectures are not equipped to represent these temporal relations.

The framework for the mechanism of shared world knowledge required the reconceptualization of several additional concepts, in addition to the previously modeled **beliefs**, **actions** and **desires**. In subsection 3.2.4 we discussed how we could model the additional concepts of **knowledge** and **intentions** for RL-based architectures. However, we notice that none of the architectures discussed in the literature review (see section 2.2) are currently equipped to generate **hypotheses** related to **beliefs**, **desires** and **intentions**. Furthermore, only two architectures are able to choose the correct **hypothesis** from a selection of multiple **hypotheses**. Thus we can conclude that none of the most commonly found architectures are able to reason about this mechanism of shared world knowledge, on account of these architectures being unable to generate **hypotheses** related to the concepts of **beliefs**, **desires** and **intentions**.

3.3 Experiment design

In the previous section (see section 3.2) we discussed how we could reconceptualize the frameworks for the mechanisms of visual perspective-taking, interpreting actions and shared world knowledge. We discussed how we could reconceptualize several concepts, like **beliefs**, **desires**, **intentions** and **actions**, for existing RL-based architectures. With these reconceptualizations of the frameworks defined in Figure 3.2 and 3.5, we can design tasks to evaluate existing RL-based architectures. This section will specify several tasks that can be used to evaluate RL-based architectures on the mechanisms of visual perspective-taking, interpreting actions and shared world knowledge. In the next chapter (see chapter 4), we will run these tasks on an existing implementation [44] of Rabinowitz et al.'s paper [77], as a use case. We start by defining components that can be used in all following experiments:

3.3.1 Base components

We will base all of the following experiments on the classic Sally-Anne task (see subsection 3.1.2). Although this task is related to the mechanism of interpreting actions, we will show how this task can be modified to also test the following concepts: The mechanism of visual perspective-taking, modeling knowledge, and modeling the relation between desire and intention. As stated in subsection 3.2.4, we are currently unable to model tasks related to the mechanism of shared world knowledge for the most commonly used aToM architectures, on account of these models being unable to generate and then choose a correct hypothesis. Therefore, no experiments will be included in order to test these specific concepts. In order to explain how the classic Sally-Anne task is modified for a specific experiment, a text-based task description is added to each experiment description. The experiments listed below will make use of the following components:

- An **agent** representing **Sally**.
- A number of **boxes**, with two properties:
  - Transparancy: The box can either be transparant, i.e. made of glass, or opaque, ensuring Sally can't see the contents of the box.
  - A marker, telling the model (not the agent) whether a box currently contains a marble, whether the box originally contained said marble, or has been empty all throughout the experiment.
- a (number of) **marble(s)**. If multiple marbles are present in an experiment, they can be distinguished based on color. Thus each marble has a unique color.

All of the following experiments make use of an **environment**, represented as an MDP. The environment should consist of a number of **states**, one state for each box in the experiment, and an extra state representing the starting position of the **agent** (Sally). The **boxes** can be represented

as a **feature** of a specific state. Other information, such as which **marble** is in which **box**, can also be represented as a **feature** of a specific state. The state containing Sally should have a number of **actions**, each action representing looking in a specific box, and **transitioning** (moving) the agent's state to the state that contains that specific box.

In the experiment descriptions, we will represent this environment as a **3x3 gridworld**. The agent is positioned at the center of this gridworld, and is limited to a single action, choosing from {moving up, moving down, moving left, moving right}. Surrounding the agent are 4 boxes. Each action then represents looking in one of the four boxes.. However, the specifics of how this environment should be represented should depend on the specifics of the model that we want to test on. In each experiment description we will explain what the goals of each specific experiment are, how the training set and test environment should be designed, and what the experiment represents in terms of a human ToM task.

Some of the experiments below have a random component. It is always assumed that any aspect of the experiment that is chosen randomly, is drawn from a uniform distribution.

3.3.2 Control experiment

In order to validate the functioning of an aToM model using the base components (see section 3.3.1), the following control experiment is proposed:

- **Training and test environment:** A 3x3 gridworld with 4 boxes.
- **Agent behavior:** The agent chooses one of these boxes at random.

The expected accuracy for the model using this specific configuration, is to be accurate 25% of the time. This is because, if the model learns to model a uniform distribution, it will have a 1 in 4 chance of choosing the correct box to look in for a specific test sample.

A human equivalent of this experiment would be the following task:

"A room contains four boxes. Daniel looks into one of the boxes. In which box does Daniel look?" Answer: He chooses one at random.

3.3.3 Visual perspective-taking experiment

In order to validate whether the artificial ToM model can reason about the mechanism of visual perspective-taking, the following experiment is proposed:

- **Training and test environment**: A 3x3 gridworld with 4 glass boxes. One box, chosen randomly, contains a marble.

- **Agent behavior:** The agent moves to the marble.

A human equivalent of this experiment would be the following task[4]:

"The room contains 4 glass boxes, one of which contains a marble. Sophie stands in the middle of these 4 boxes. Point in the direction of the box with the marble, from Sophie's perspective."

As stated in subsection 3.2.2, RL-based architectures can explicitly mode spatial relations. Therefore we can use the reconceptualization of the framework for visual perspective-taking from subsection 3.2.2 to design this experiment: We can model **task-based beliefs** as the positions of the 4 glass boxes, one containing a marble. We can explicitly model the **actions**: {pointing up, pointing left, pointing down, pointing right}. We can model Sophie's **desire** as: Sophie points to the marble. This is learned during training. Thus this experiment can be used to test the mechanism of visual perspective-taking for RL-based aToM models.

3.3.4 Interpreting actions experiment

We can test the mechanism of interpreting actions for aToM models, in a similar fashion to the mechanism of visual perspective-taking:

- **Training and test environment**: A 3x3 gridworld with 4 boxes. One box, chosen randomly, contains a marble. Another box, chosen randomly, is marked as the box where the marble was originally placed.
- **Agent behavior:** The agent moves to where the marble was originally placed.

A human equivalent of this experiment would be the following task:

"The room contains a white, black, red and blue box. Sally puts a marble in a white box, then leaves the room. Anne takes the marble from white box and puts it in a black box, then Sally comes back and looks for her marble. Where will Sally look for her marble?" Answer: In the white box, because she believes her marble to be there.

This experiment represents the classic Sally Anne experiment, with additional boxes. These additional boxes serve as a control (as they never contained the marble), and exist to diversify the training set. Now the agent can't simply learn to repeat a single learned scenario (as there are a total of 12 different scenarios), but has to learn to separate the agents beliefs from their own beliefs in order to solve the task at hand. As stated in subsection 3.2.3, RL-based architectures are not equipped to represent temporal relations, so therefore we have to mark the box in which the marble

[4]It should be noted that tasks related to the mechanism of visual perspective-taking are usually presented in a visual format. Therefore a more accurate human equivalent of this experiment would be an image containing the four boxes, as well as Sophie's position as outlined in this task description.

was originally placed, rather than first placing the marble in one box, and then moving it to another box.

In order to be able to even better test the ability of the model to generalize, the training set can consist of less than all 12 of these different scenarios. For instance, let the training set never contain the specific scenario in which Sally placed a marble in the leftmost box, then Anne moved it to the rightmost box. The test set could then contain this specific scenario. It would be interesting to evaluate how many of the twelve scenarios can be omitted, while still achieving a high accuracy on the to be tested model.

3.3.5 Learning knowledge experiment

As stated in subsection 3.2.4, none of the architectures discussed in the literature review (see section 2.2), are equipped to model all components found in the shared world knowledge framework from Figure 3.5. They are unable to generate multiple hypotheses and then subsequently reason which of these hypotheses will lead to solving a ToM task related to the mechanism of shared world knowledge. However, some of the other components, specifically learning knowledge, and modeling desires and intentions, can still be tested for current aToM architectures. The following experiment will test the ability of a current RL-based aToM model to learn knowledge:

- **Training environment**: A 3x3 gridworld containing 4 boxes, 0 to 4 of which are made of glass. One box, chosen randomly, contains a marble. Another box, chosen randomly, is marked as the location where the marble is placed.
- **Test environment**: A 3x3 gridworld containing 4 boxes, 2 of which are made of glass. One box, chosen randomly, contains a marble. Another box, chosen randomly, is marked as the box where the marble was originally placed.
- **Agent behavior:** The agent will move to the marble if it is placed in a glass box. The agent will move to where the marble was placed if both the marble, and the original location of the marble were non-glass boxes. If the marble was placed in a glass box, but is currently not in a glass box, the agent will choose a non-glass box at random.

The goal of this experiment is to test whether the aToM model can learn knowledge from a set of example scenarios, as explained in section 3.2.4 - Modeling knowledge. The learned knowledge in this experiment consists of learning how non-glass and glass boxes respectively hide and reveal their contents. The testing scenario of 2 glass boxes was chosen specifically, because it represents the most complicated scenario:

"A room has 4 boxes. Box 1 and 2 are not made of glass (and are opaque). Box 3 and 4 are made of glass. Sam places his marble in box 3, and then leaves the room, now Andy moves the marble from box 3 to box 1. Then Sam comes back and looks for his marble. In what box will Sam look for his marble?" Answer: Sam will choose either box 1 or 2, because he can see that his marble is not in box 3 or 4.

Like with the 'interpreting actions' experiment in section 3.3.4, specific configurations can be omitted from the training set, in order to test under what minimal circumstances the to be tested artificial ToM model still produces accurate results.

3.3.6 Desire and intentions experiment

Although all other experiments already include reasoning about desire and intentions, we can model these components in a more explicit manner, using the following experiment:

- **Training environment**: A 3x3 gridworld containing 2 marbles, chosen at random from the color {red, blue, green, yellow}.
- **Test environment**: A 3x3 gridworld containing all four colors marbles.
- **Agent behavior:** The agent will choose the marble as follows: red>blue>green>yellow.

The goal of this experiment is to test whether the artificial ToM model can learn to reason about an actor's intention, from a learned desire. The agent could have differing intentions, depending on which color of marble they prefer. The model should learn the desire of the agent from the examples provided in the training set, and then reason about the agent's intention (following from their desire) in the test set. Instead of the desire (red>blue>green>yellow) provided in the experiment description, the relative preferences for certain colors of marbles could be altered (for instance: yellow>green>red>blue). A human equivalent of this experiment would be the following task:

"Olivia prefers red marbles to blue marbles. She also prefers green marbles to yellow marbles. Lastly, she prefers blue marbles to green marbles. She is presented with a choice of 4 marbles, one red, one blue, one green and one yellow. Which marble will she choose?" Answer: The red marble, because she prefers it over all the other colors of marble.

Chapter 4 - Use case

4.1 Model specification

In order to show how the tasks from section 3.3 can be applied to an existing aToM model, all tasks were applied to an implementation [44] of Rabinowitz et al.'s model [77]. For more information on Rabinowitz's model, see subsection 2.2.5. This specific implementation of the model uses an 11x11 gridworld, with a third dimension being used in order to encode state features in the gridworld, such as the placement of the agent or boxes. Each feature is given its own layer. For example, the fifth layer of a state encodes for "contains a box". We use binary encoding to encode all features. For our example, 1 means: a glass box exists at this state. 0 means: no glass box exists at this state. If the state at coordinate (1,1) would have a box, and the state at coordinate (2,1) would not have a box, this would mean that (1,1,5) = 1 (because it contains a box), and (2,1,5) = 0 (because it does not contain a box). All features needed for the experiments were encoded in this manner. The dimensions of the gridworld, as well as the number of allowed features, were hard-coded into the structure of the model. In order to modify the model as minimally as possible, the 11x11 dimensions of the gridworld were kept. However, the model was modified to allow for a variable number of features, depending on the experiment.

Each of the experiments were implemented as described in section 3.3. The 3x3 gridworld described in each experiment description was placed at the center of the 11x11 gridworld, but otherwise the description (see 'training environment', 'test environment' and 'agent behavior' for each experiment) remains the same. A unique feature layer was reserved for each of the following base components, defined in subsection 3.3.1:

- Agent
- Opaque boxes
- Transparent boxes
- One layer per color of marble
- One layer per color of marble, to mark in which box those marbles were placed

Only the minimum number of feature layers were used, based on what was required for a specific experiment. For instance, if an experiment did not require the use of transparent boxes, no feature layer was included to represent transparent boxes. The model was trained using 1000 training samples, and tested using a different 1000 samples.

4.2 Results

The results of each experiment are listed below (see Table 4.1):

Experiment	**Training acc%**	**Testing acc%**
Control	25%	25%
Visual perspective-taking	100%	100%
Interpreting actions	100%	100%
Learning knowledge	94%	83%
Desire and intentions	100%	100%

Table 4.1: Results of the pilot experiment. For the experiment descriptions, see section 3.3. We find a maximum possible accuracy on all experiments for both the training and test set.

The accuracy in all tests is the maximum possible accuracy that can be achieved. The 25% accuracy in the control experiment corresponds to the model choosing one of the four boxes at random. The 94% and 83% accuracy results for the knowledge experiment are caused by the fact that the actor will randomly choose a non-glass box if the marble was originally in a glass box, but is now in an opaque box. For the test set (2 glass boxes), we can list out all the possible configurations (two out of four glass boxes, marble placed in one at random, and originally placed in another), and calculate the possibly achievable accuracy for each scenario (see Table 4.2). If we do this, we find that $\frac{2}{3}$ of the scenarios produce non-random behavior, and are reproducible with a 100% accuracy. In the other $\frac{1}{3}$ scenarios, the agent will choose one of the two non-glass boxes at random, leading to a maximum achievable accuracy of 50% for the model. Therefore, the maximum achievable accuracy for all these scenarios combined will give an accuracy of $\frac{1}{3} * 50 + \frac{2}{3} * 100 = 83.3333\%$. We can calculate this result similarly for 0, 1, 3 and 4 glass boxes, leading to an accuracy of 94,343333% for the training set.

Marble placed/moved to	**Max acc%**	**Num configurations**	**Behavior**
Glass/Glass	100%	12	Picks box containing the marble
Glass/Opaque	50%	24	Picks one of the two opaque boxes at random
Opaque/Glass	100%	24	Picks box containing the marble
Opaque/Opaque	100%	12	Picks box in which the marble was originally placed
Average:	83%		

Table 4.2: Maximum achievable accuracy for all possible configurations consisting of four boxes, two of which are made of glass. One marble is placed in one of the four boxes, and moved to another box. The actor's behavior is specified for each subconfiguration. The average max acc% is the weighted average, obtained by multiplying the max acc% with the number of configurations for each subconfiguration.

Chapter 5 - Conclusion and Discussion

In this thesis, we analyzed tasks and mechanisms related to Theory of Mind. Specifically, we examined what mechanisms comprise ToM in humans, and how these mechanisms are measured. We then formalized these mechanisms, in order to analyze how artificial ToM models could reason about them. We also designed tasks to evaluate aToM models on key concepts related to ToM. This was done in order to address three key problems relating to the creation of artificial Theory of Mind (see chapter 1): 1) ToM is an ambiguous and multifaceted concept. What mechanisms comprise ToM? 2) ToM can only be measured indirectly for humans, so how do we measure ToM skills in artificial models? 3) It is currently unknown what is required of aToM models in order to reason succesfully about ToM [7]. In order to address these key problems, we formulated the following research questions:

RQ1: How can one formally define the different mechanisms relating to ToM, and evaluate aToM models' effectiveness in reasoning about said mechanisms?

RQ1.1: What mechanisms are relevant for state-of-the-art aToM research?

RQ1.2: What are the requirements for aToM models to be able to effectively reason about these mechanisms?

RQ1.3: How can these mechanisms be evaluated for artificial ToM models?

In this chapter we will address our findings, answering these questions. We will also discuss the potential implications and pitfalls of this research, as well as provide recommendations for future work. We will start by discussing subquestions RQ1.1, RQ1.2 and RQ1.3.

5.1 RQ1.1: Relevant aToM mechanisms

As stated in chapter 1, ToM is a highly ambiguous concept, even when human ToM research is concerned. In order to find what mechanisms humans use when reasoning using ToM, a more precise definition of what constitutes ToM was required. Quesque and Rossetti [76] defined two criteria, against which tasks can be evaluated in order to analyze whether ToM is required in order to solve them (see subsection 2.1.2). These criteria are:

1. **Nonmerging criterion:** Participants need to make a distinction between their own mental state, and the mental state they infer.

2. **Mentalizing criterion:** Lower-level processes should not account for successful performance on ToM tasks.

Using these criteria we listed several of the most common tasks (based on Quesque and Rossetti's research [76]) used in ToM research that satisfy these two criteria (see subsection 2.1.3). Consecutively, these tasks were grouped based on similarity into three categorizations representing the mechanisms humans employ in order to solve them. This categorization was based on research by Byom and Mutlu [17]. We defined three categorizations: Interpreting actions, visual perspective-taking and shared world knowledge (see subsection 2.1.4).

In section 3.1, we dissected sample tasks for each categorization, in order to examine what mechanisms humans employ in order to solve them. We created a framework for how humans solve tasks related to each categorization, representing the mechanism through which they reason about said tasks. We also looked at what auxiliary skills are required in order to reason about tasks in each category. As stated before, ToM is a multifaceted concept, that can only be measured indirectly. Therefore, additional skills are required in order to solve tasks relating to ToM. We defined auxiliary skills as skills that, although required in order to solve the task, do not pass the nonmerging or mentalizing criteria (see subsection 2.1.2).

The frameworks for each mechanism are based on the BDI model (see 3.1.1). The BDI model is a widely used model to explain reasoning related to ToM, and is rooted in theory theories [5], one of the leading theories on how humans reason about ToM. Theory theories propose that humans reason about ToM using a set of concepts, such as beliefs, desires and intentions (the concepts from the BDI model), and the relation between these concepts. We chose to base the frameworks on the BDI model, as during the literature review (see section 2.2) it was found that many of the most commonly used aToM architectures make use of this BDI model. However, using the BDI model as a basis for formalizing ToM mechanisms can be seen as a limitation of this study, as currently, there exists no consensus on how humans reason about ToM (a problem mentioned in chapter 1). For instance, an adverse theory to theory theories, called simulation theories [5], states that humans reason about ToM by simulating themselves from the perspective of another person. Therefore, the created frameworks can be considered to be an interpretation of ToM mechanisms based on theory theories.

It was found, that tasks categorized under the mechanisms of interpreting actions and visual perspective-taking required the participant to reason about ToM in a similar fashion (see subsections 3.1.2 and 3.1.3). Both mechanisms require the participant to separate the actor's beliefs from their own beliefs, in order to solve tasks related to these mechanisms. For an overview of the framework for these mechanisms, see Figure 3.2. Thus we can consider tasks categorized under either mechanism (interpreting actions or visual perspective-taking) to require the same ToM mechanism in order to solve them. However, tasks categorized under these mechanisms differ in the auxiliary skills required in order to solve them. Tasks categorized under interpreting actions are always presented either in written format or being told by a researcher, thus requiring the participant to have language comprehension skills. For tasks categorized under visual perspective-taking, spatial orientation skills are required, as all of these tasks require the participant to understand spatial relations.

Tasks categorized under the mechanism of shared world knowledge were found to require a different

ToM mechanism in order to solve them (see subsection 3.1.4). The framework for how humans reason about tasks related to this mechanism can be found in Figure 3.5. In order to solve tasks related to this mechanism, participants have to generate hypotheses related to beliefs, desire and intentions to explain the actor's behavior, and then choose the most likely hypothesis based on their own beliefs.

5.2 RQ1.2: Requirements for aToM models

For the purpose of finding what is required of aToM models in order to reason about mechanisms relating to ToM, we examined how the mechanisms found in section 3.1, could be reconceptualized for the most commonly found model architectures used in current aToM research. To find what architectures are commonly used, we conducted a literature review on papers that researched aToM, and created an aToM model (see section 2.2). We found five categories of model architectures used in current aToM research. Two of these categories are rule-based (game theory-based architectures and cognitive architectures), executing behavior explicitly defined by a researcher. Three of these categories include a learning component, and are based on machine learning principles (observational RL, inverse RL and Bayesian inference-based architectures). These architectures had some commonalities: Many of the found architectures make use of the BDI model as a basis for modeling the concepts of beliefs, desires and intentions, and all machine learning based architectures make use of a simulated environment using Markov Decision Processes (MDPs) [93], modeling states, possible actions for each state, and a transition function in order to transition between states using actions.

In section 3.2, we used these found commonalities to reconceptualize the frameworks presented in section 3.1 for the models that generalize well (machine learning-based architectures). We discussed how to model the concepts defined in the BDI model for models using MDPs (see subsection 3.2.2), noting for which concepts their generalization is robust, and for which concepts it is not. We also presented several recommendations for the creation of future aToM models, addressing areas in which we noted poor generalization for the current models.

It was found that the framework for the mechanism of visual perspective-taking (see Figure 3.2) could be reconceptualized for aToM models using MDPs. This is, because tasks related to visual perspective-taking are presented in a visual format, requiring the participant to represent spatial relations. MDPs can be represented as a gridworld [93]. Using this gridworld, one can represent spatial relations. We could also model the concepts required for the framework for visual perspective-taking (see subsection 3.2.2) for aToM models using MDPs. However, we do note poor generalization to the concept of desire, requiring a model to be retrained every time an actor has a different desire. It is recommended that future aToM models explicitly model this concept of desire as a parameter, allowing for better generalization towards this concept. On the other hand, rule-based cognitive architectures allow researchers to explicitly define multiple desires. However, if one is to design a future rule-based cognitive architecture, based on the BDI model, it is recommended to explicitly model spatial relations, as none of the cognitive architectures analyzed during the literature review (see subsection 2.2.3) allow for representation of these spatial relations.

As the mechanism of interpreting actions requires similar ToM reasoning to the mechanism of visual perspective-taking (using the same framework, see Figure 3.2), the same manner of reconceptualizing applies to this mechanism (see subsection 3.2.3). We again note a poor generalization towards the concept of desire. Furthermore, it was found that tasks categorized under the mechanism of interpreting actions include temporal relations. None of the aToM models using MDPs allow us to model temporal relations in the environment, other than the agent. It is therefore recommended that future aToM models include some manner to to model these temporal relations, for instance by allowing objects in the environment, other than the agent, to be moved.

The mechanism of shared world knowledge requires several additional concepts to be modeled (see subsection 3.2.4). Some of these additions can be modeled with relative ease. However, none of the models discussed in the literature review (see section 2.2) allow for the generation, and subsequent testing of hypotheses. Humans have the ability to quickly disregard hypotheses that are either irrelevant, or are obviously true [43]. However, all of the models discussed need a complete representation of all possible beliefs, desires and intentions. In this sense, none of the found models are yet suited to reasoning about real-world ToM situations, rather than simple ToM tasks. For instance, consider a real-world enactment of the Sally-Anne task [12], with a human participant. Sally puts the marble in a box and leaves the room. However, when Sally comes back into the room, she is now crying. If a human participant was present in the room during this task, they would likely be inclined to ask why Sally is crying, or try to comfort her. They would not assume that Sally will look into either of the boxes at that point, as something (apparently) has come up. However, none of the models discussed thus far can generalize to this unanticipated outcome, unless it was specifically modeled beforehand as a possibility. This shows that none of the models discussed during the literature review (see section 2.2) can currently generalize well enough to be applied to real-world scenarios. It appears that this efficient generation of hypotheses can be considered to be one of the biggest obstacles in creating a human level aToM model.

5.3 RQ1.3: Evaluating aToM models

In order to evaluate existing aToM models on the mechanisms of visual perspective-taking and interpreting actions, as well as concepts related to the mechanism of shared world knowledge, several experiments were designed (see section 3.3) and subsequently tested on an existing implementation [44] of Rabinowitz et al.'s model [77] (see chapter 4). The goal of these experiments was to provide a proof-of-concept, showing how we could reconceptualize the concepts defined in the frameworks from Figures 3.2 and 3.5, and evaluate them on an existing model. The model on which we ran the experiments achieved a maximum possible accuracy on all possible experiments, showing that this specific model is able to reason about the concepts related to visual perspective-taking and interpreting actions, as well as being able to learn knowledge and learn how desires and intentions relate to each other.

However, it should be noted that these problems only require the model to use ToM reasoning in a limited setting, requiring a model to learn only a small amount of possible configurations in order to solve the task with maximum accuracy. This is similar to how the ToM tasks for human participants

(see subsection 2.1.3) only test ToM skills in a limited environment. Although these tasks allowed us to formalize what mechanisms humans use in order to employ ToM (see section 3.1), real-world problems require humans to use ToM in a much more complicated setting.

5.4 Future work

The research presented in this thesis provides opportunities for the creation of future aToM models. Incorporation of the recommendations (see section 3.2) presented in this thesis should allow these future models to be better able to reason about ToM using the mechanisms presented in this thesis. Specifically, effective representation of the generation and subsequent testing of hypotheses should allow future aToM models to reason about more complex ToM-related tasks, allowing these models to more accurately reason about real-world problems requiring ToM.

Besides the creation of future aToM models, more research can be done on how the mechanisms presented in this thesis relate to real-world usage of ToM. The mechanisms presented in this thesis can be seen as the building blocks for solving real-world ToM problems. However, the research done in this thesis does not reveal what other skills are required in order to develop artificial ToM at a near-human level.

5.5 Conclusion

This thesis presented research on what mechanisms constitute ToM, as well as how we can represent these mechanisms in an artificial context. We found three categorizations of ToM tasks, and formulated how aToM models can reason about each categorization of task. We also designed experiments to test whether current models can effectively reason about each type of task. It was found that aToM models can reason about tasks related to the mechanisms of visual-perspective taking, and interpreting actions. However, currently, the most commonly found aToM models fail to reason about the mechanism of shared world knowledge, on account of these models being unable to generate and subsequently test hypotheses. This inability to efficiently generate and reason about hypotheses limits these models' effectiveness when reasoning about more complex real-world problems requiring the use of ToM, and can be considered to be one of the biggest obstacles for the creation human-level aToM models.

Bibliography

[1] Elsevier Scopus, 2004.

[2] Regulatory framework proposal on artificial intelligence | Shaping Europe's digital future, Sept. 2023.

[3] Ajzen, I. The theory of planned behavior. *Organizational Behavior and Human Decision Processes 50*, 2 (Dec. 1991), 179–211.

[4] Akula, A. R., Liu, C., Saba-Sadiya, S., Lu, H., Todorovic, S., Chai, J. Y., and Zhu, S.-C. X-ToM: Explaining with Theory-of-Mind for Gaining Justified Human Trust, Sept. 2019. arXiv:1909.06907 [cs].

[5] Apperly, I. A. Beyond Simulation–Theory and Theory–Theory: Why social cognitive neuroscience should use its own concepts to study "theory of mind". *Cognition 107*, 1 (Apr. 2008), 266–283.

[6] Apperly, I. A. *Mindreaders | The Cognitive Basis of "Theory of Mind"*, 1st ed. Psychology Press, Oct. 2010.

[7] Aru, J., Labash, A., Corcoll, O., and Vicente, R. Mind the gap: Challenges of deep learning approaches to Theory of Mind, Mar. 2022. arXiv:2203.16540 [cs, q-bio].

[8] Astington, J. W., and Jenkins, J. M. Theory of mind development and social understanding. *Cognition and Emotion 9*, 2-3 (Mar. 1995), 151–165. Publisher: Routledge _eprint: https://doi.org/10.1080/02699939508409006.

[9] Baker, C. L., Jara-Ettinger, J., Saxe, R., and Tenenbaum, J. B. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour 1*, 4 (Mar. 2017), 0064.

[10] Baker, C. L., Saxe, R. R., and Tenenbaum, J. B. Bayesian Theory of Mind: Modeling Joint Belief-Desire Attribution. *Proceedings of the annual meeting of the cognitive science society 33* (2011).

[11] Bara, C.-P., CH-Wang, S., and Chai, J. MindCraft: Theory of Mind Modeling for Situated Dialogue in Collaborative Tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (2021), pp. 1112–1125. arXiv:2109.06275 [cs].

[12] Baron-Cohen, S., Leslie, A. M., and Frith, U. Does the autistic child have a "theory of mind" ? *Cognition 21*, 1 (Oct. 1985), 37–46.

[13] Baron-Cohen, S., O'Riordan, M., Stone, V., Jones, R., and Plaisted, K. Recognition of faux pas by normally developing children and children with Asperger syndrome or high-functioning autism. *Journal of Autism and Developmental Disorders 29*, 5 (Oct. 1999), 407–418.

[14] Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., and Plumb, I. The "Reading the Mind in the Eyes" Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry, and Allied Disciplines 42*, 2 (Feb. 2001), 241–251.

[15] Bosse, T., Memon, Z. A., and Treur, J. A Two-Level BDI-Agent Model for Theory of Mind and its Use in Social Manipulation. *Proceedings of the AISB 2007 Workshop on Mindful Environments 4* (2007).

[16] Bratman, M. *Intention, Plans, and Practical Reason.* Cambridge, MA: Harvard University Press, Cambridge, 1987.

[17] Byom, L., and Mutlu, B. Theory of mind: mechanisms, methods, and new directions. *Frontiers in Human Neuroscience 7* (2013).

[18] Byrne, M. D. COGNITIVE ARCHITECTURE. In *The Human-Computer Interaction Handbook*, 2 ed. CRC Press, 2007. Num Pages: 22.

[19] Cominelli, L., Mazzei, D., and De Rossi, D. E. SEAI: Social Emotional Artificial Intelligence Based on Damasio's Theory of Mind. *Frontiers in robotics and AI 5* (2018), 6.

[20] Cuff, B., Brown, S., Taylor, L., and Howat, D. Empathy: A Review of the Concept. *Emotion Review 8* (Apr. 2016), 144–153.

[21] Cuzzolin, F., Morelli, A., Cîrstea, B., and Sahakian, B. J. Knowing me, knowing you: theory of mind in AI. *Psychological Medicine 50*, 7 (May 2020), 1057–1061. Publisher: Cambridge University Press.

[22] Domingos, P. The Role of Occam's Razor in Knowledge Discovery. *Data Mining and Knowledge Discovery 3*, 4 (Dec. 1999), 409–425.

[23] Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., Kessler, J., Woike, J. K., Wolf, O. T., and Convit, A. Introducing MASC: A Movie for the Assessment of Social Cognition. *Journal of Autism and Developmental Disorders 36*, 5 (July 2006), 623–636.

[24] Eliasmith, C. Attractive and In-discrete. *Minds and Machines 11* (2001).

[25] Epley, N., and Caruso, E. M. Perspective taking: Misstepping into others' shoes. In *Handbook of imagination and mental simulation.* Psychology Press, New York, NY, US, 2009, pp. 295–309.

[26] Erle, T., and Topolinski, S. Spatial and Empathic Perspective-Taking Correlate on a Dispositional Level. *Social Cognition 33* (June 2015), 187–210.

[27] Erle, T. M., and Topolinski, S. The grounded nature of psychological perspective-taking. *Journal of Personality and Social Psychology 112*, 5 (2017), 683–695. Place: US Publisher: American Psychological Association.

[28] Flavell, J. H., Flavell, E. R., and Green, F. L. Development of the appearance-reality distinction. *Cognitive Psychology 15*, 1 (Jan. 1983), 95–120.

[29] Friedlander, D., and Franklin, S. LIDA and a Theory of Mind. *Frontiers in Artificial Intelligence and Applications 171* (2008).

[30] Frith, C. D., and Frith, U. The neural basis of mentalizing. *Neuron 50*, 4 (May 2006), 531–534.

[31] Gallese, V., and Goldman, A. Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences 2*, 12 (Dec. 1998), 493–501.

[32] Giovagnoli, A. R., Bell, B., Erbetta, A., Paterlini, C., and Bugiani, O. Analyzing theory of mind impairment in patients with behavioral variant frontotemporal dementia. *Neurological Sciences 40*, 9 (Sept. 2019), 1893–1900.

[33] González, B., and Chang, L. J. Computational Models of Mentalizing. In *The Neural Basis of Mentalizing*, M. Gilead and K. N. Ochsner, Eds. Springer International Publishing, Cham, 2021, pp. 299–315.

[34] Goodman, N. D., Tenenbaum, J. B., Feldman, J., and Griffiths, T. L. A Rational Analysis of Rule-Based Concept Learning. *Cognitive Science 32*, 1 (2008), 108–154. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1080/03640210701802071.

[35] Gopnik, A., and Meltzoff, A. N. *Words, thoughts, and theories.* Words, thoughts, and theories. The MIT Press, Cambridge, MA, US, 1997. Pages: xvi, 268.

[36] Hale, C. M., and Tager-Flusberg, H. The influence of language on theory of mind: a training study. *Developmental Science 6*, 3 (2003), 346–359. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-7687.00289.

[37] Hamilton, A. F. d. C., Brindley, R., and Frith, U. Visual perspective taking impairment in children with autistic spectrum disorder. *Cognition 113*, 1 (2009), 37–44. Place: Netherlands Publisher: Elsevier Science.

[38] Happé, F. G. E. An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders 24*, 2 (Apr. 1994), 129–154.

[39] Hedden, T., and Zhang, J. What do you think I think you think?: Strategic reasoning in matrix games. *Cognition 85*, 1 (Aug. 2002), 1–36.

[40] Hegarty, M., and Waller, D. A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence 32*, 2 (Mar. 2004), 175–191.

[41] Heider, F., and Simmel, M. An Experimental Study of Apparent Behavior. *The American Journal of Psychology 57*, 2 (1944), 243–259. Publisher: University of Illinois Press.

[42] Hiatt, L., Harrison, A., and Trafton, J. *Accommodating Human Variability in Human-Robot Teams through Theory of Mind.* Jan. 2011. Journal Abbreviation: IJCAI International Joint Conference on Artificial Intelligence Pages: 2071 Publication Title: IJCAI International Joint Conference on Artificial Intelligence.

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

[44] Jeon, K. The Implementation of "Machine Theory of Mind", ICML 2018, Mar. 2023. original-date: 2021-11-29T06:40:25Z.

[45] Kanske, P., Böckler, A., Trautwein, F.-M., Parianen Lesemann, F. H., and Singer, T. Are strong empathizers better mentalizers?: Evidence for independence and interaction between the routes of social cognition. *Social Cognitive and Affective Neuroscience 11*, 9 (Sept. 2016), 1382–1392. Publisher: Oxford University Press.

[46] Keysar, B. The Illusory Transparency of Intention: Linguistic Perspective Taking in Text. *Cognitive Psychology 26*, 2 (Apr. 1994), 165–208.

[47] Keysar, B., Ginzel, L. E., and Bazerman, M. H. States of Affairs and States of Mind: The Effect of Knowledge of Beliefs. *Organizational Behavior and Human Decision Processes 64*, 3 (Dec. 1995), 283–293.

[48] Kim, K.-J., and Lipson, H. Towards a "theory of mind" in simulated robots. In *Proceedings of the 11th Annual Conference Companion on Genetic and Evolutionary Computation Conference: Late Breaking Papers* (Montreal Québec Canada, July 2009), ACM, pp. 2071–2076.

[49] Klatt, J., Marsella, S., and Krämer, N. *Negotiations in the Context of AIDS Prevention: An Agent-Based Model Using Theory of Mind*. Sept. 2011. Pages: 215.

[50] Kosinski, M. Theory of Mind May Have Spontaneously Emerged in Large Language Models, Mar. 2023. arXiv:2302.02083 [cs].

[51] Lai, M.-C., Lombardo, M. V., and Baron-Cohen, S. Autism. *The Lancet 383*, 9920 (Mar. 2014), 896–910.

[52] Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. Building machines that learn and think like people. *Behavioral and Brain Sciences 40* (Jan. 2017), e253. Publisher: Cambridge University Press.

[53] Langley, C., Cirstea, B. I., Cuzzolin, F., and Sahakian, B. J. Theory of Mind and Preference Learning at the Interface of Cognitive Science, Neuroscience, and AI: A Review. *Frontiers in artificial intelligence 5* (Jan. 2022), 778852.

[54] Legerstee, M., Barna, J., and DiAdamo, C. Precursors to the development of intention at 6 months: Understanding people and their actions. *Developmental Psychology 36* (2000), 627–634. Place: US Publisher: American Psychological Association.

[55] Marchetti, A., Di Dio, C., Cangelosi, A., Manzi, F., and Massaro, D. *Developing ChatGPT's Theory of Mind*. Mar. 2023.

[56] McDonald, P. S., Bornhofen, C., Shum, D., Long, E., Saunders, C., and Neulinger, K. Reliability and validity of The Awareness of Social Inference Test (TASIT): A clinical test of social perception. *Disability and Rehabilitation* (July 2009). Publisher: Taylor & Francis.

[57] Melhart, D., Yannakakis, G. N., and Liapis, A. I Feel I Feel You: A Theory of Mind Experiment in Games. *KI - Künstliche Intelligenz 34*, 1 (Mar. 2020), 45–55.

[58] Misuraca, G., van Noordt, C., and Boukli, A. *The use of AI in public services: results from a preliminary mapping across the EU*. Sept. 2020.

[59] Ng, A. Y., and Russell, S. J. Algorithms for Inverse Reinforcement Learning. In *Proceedings of the Seventeenth International Conference on Machine Learning* (San Francisco, CA, USA, June 2000), ICML '00, Morgan Kaufmann Publishers Inc., pp. 663–670.

[60] Nguyen, T. N., and Gonzalez, C. Theory of Mind From Observation in Cognitive Models and Humans. *Topics in Cognitive Science 14*, 4 (2022), 665–686. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tops.12553.

[61] Noble, W. S. What is a support vector machine? *Nature Biotechnology 24*, 12 (Dec. 2006), 1565–1567. Number: 12 Publisher: Nature Publishing Group.

[62] Oguntola, I., Hughes, D., and Sycara, K. Deep Interpretable Models of Theory of Mind. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)* (Aug. 2021), pp. 657–664. ISSN: 1944-9437.

[63] Ono, T., Imai, M., and Nakatsu, R. Reading a robot's mind: a model of utterance understanding based on the theory of mind mechanism. *Advanced Robotics 14*, 4 (Jan. 2000), 311–326.

[64] OpenAI. GPT-4 Technical Report, Mar. 2023. arXiv:2303.08774 [cs].

[65] Owen, G. *Game Theory*. Emerald Group Publishing, Aug. 2013. Google-Books-ID: yeVbAAAAQBAJ.

[66] Pak, M., and Kim, S. A review of deep learning in image recognition. In *2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT)* (Aug. 2017), pp. 1–3.

[67] Panisson, A. R., Sarkadi, S., McBurney, P., Parsons, S., and Bordini, R. H. On the Formal Semantics of Theory of Mind in Agent Communication. In *Agreement Technologies*, M. Lujak, Ed., vol. 11327. Springer International Publishing, Cham, 2019, pp. 18–32. Series Title: Lecture Notes in Computer Science.

[68] Papagni, G., and Koeszegi, S. Understandable and trustworthy explainable robots: A sensemaking perspective. *Paladyn, Journal of Behavioral Robotics 12* (Oct. 2020), 13–30.

[69] Patacchiola, M., and Cangelosi, A. A Developmental Cognitive Architecture for Trust and Theory of Mind in Humanoid Robots. *IEEE Transactions on Cybernetics 52*, 3 (Mar. 2022), 1947–1959.

[70] Perner, J., and Wimmer, H. "John thinks that Mary thinks that..." attribution of second-order beliefs by 5- to 10-year-old children. *Journal of Experimental Child Psychology 39*, 3 (June 1985), 437–471.

[71] Piaget, J. *Child's Conception of Space: Selected Works vol 4*. Routledge, Aug. 2013. Google-Books-ID: iZeAAAAAQBAJ.

[72] Premack, D., and Woodruff, G. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences 1*, 4 (Dec. 1978), 515–526. Publisher: Cambridge University Press.

[73] Pylyshyn, Z. W. When is attribution of beliefs justified? [P&W]. *Behavioral and Brain Sciences 1*, 4 (Dec. 1978), 592–593. Publisher: Cambridge University Press.

[74] Pynadath, D. V., Rosenbloom, P. S., and Marsella, S. C. Reinforcement Learning for Adaptive Theory of Mind in the Sigma Cognitive Architecture. In *Artificial General Intelligence* (Cham, 2014), B. Goertzel, L. Orseau, and J. Snaider, Eds., Lecture Notes in Computer Science, Springer International Publishing, pp. 143–154.

[75] Pynadath, D. V., Wang, N., and Marsella, S. C. Are You Thinking What I'm Thinking? An Evaluation of a Simplified Theory of Mind. In *Intelligent Virtual Agents* (Berlin, Heidelberg, 2013), R. Aylett, B. Krenn, C. Pelachaud, and H. Shimodaira, Eds., Lecture Notes in Computer Science, Springer, pp. 44–57.

[76] Quesque, F., and Rossetti, Y. What Do Theory-of-Mind Tasks Actually Measure? Theory and Practice. *Perspectives on Psychological Science 15*, 2 (Mar. 2020), 384–396.

[77] Rabinowitz, N. C., Perbet, F., Song, H. F., Zhang, C., Eslami, S. M. A., and Botvinick, M. Machine Theory of Mind, Mar. 2018. arXiv:1802.07740 [cs].

[78] Richwine, L. Hollywood writers union ratifies three-year labor contract after strike. *Reuters* (Oct. 2023).

[79] Rizzolatti, G., and Craighero, L. The mirror-neuron system. *Annual Review of Neuroscience 27* (2004), 169–192.

[80] Sarfati, Y., Hardy-Bayle, M.-C., Besche, C., and Widlöcher, D. Attribution of intentions to others in people with schizophrenia: A non- verbal exploration with comic strips. *Schizophrenia research 25* (July 1997), 199–209.

[81] Sarkadi, S., Panisson, A., Bordini, R., McBurney, P., Parsons, S., and Chapman, M. Modelling deception using theory of mind in multi-agent systems. *AI Communications 32* (Aug. 2019), 1–16.

[82] Sarkadi, S., Panisson, A. R., Bordini, R. H., McBurney, P., and Parsons, S. Towards an Approach for Modelling Uncertain Theory of Mind in Multi-Agent Systems. In *Agreement Technologies* (Cham, 2019), M. Lujak, Ed., Lecture Notes in Computer Science, Springer International Publishing, pp. 3–17.

[83] Stuhlmüller, A., and Goodman, N. D. Reasoning about reasoning by nested conditioning: Modeling theory of mind with probabilistic programs. *Cognitive Systems Research 28* (June 2014), 80–99.

[84] Tanaka, H., Saga, T., Iwauchi, K., and Nakamura, S. Acceptability and Trustworthiness of Virtual Agents by Effects of Theory of Mind and Social Skills Training. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)* (Jan. 2023), pp. 1–7.

[85] Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. How to Grow a Mind: Statistics, Structure, and Abstraction. *Science 331*, 6022 (Mar. 2011), 1279–1285.

[86] Torabi, F., Warnell, G., and Stone, P. Behavioral Cloning from Observation, May 2018. arXiv:1805.01954 [cs].

[87] Veltman, K., de Weerd, H., and Verbrugge, R. Training the use of theory of mind using artificial agents. *Journal on Multimodal User Interfaces 13*, 1 (Mar. 2019), 3–18.

[88] Vinanzi, S., Patacchiola, M., Chella, A., and Cangelosi, A. Would a robot trust you? Developmental robotics model of trust and theory of mind. *Philosophical Transactions of The Royal Society B Biological Sciences 374* (Mar. 2019).

[89] Vossen, P., Baez, S., Bajčetić, L., and Kraaijeveld, B. Leolani: a reference machine with a theory of mind for social communication, June 2018. arXiv:1806.01526 [cs].

[90] Wellman, H. M. *The child's theory of mind*. The child's theory of mind. The MIT Press, Cambridge, MA, US, 1992. Pages: xiii, 358.

[91] Wellman, H. M., Cross, D., and Watson, J. Meta-Analysis of Theory-of-Mind Development: The Truth about False Belief. *Child Development 72*, 3 (2001), 655–684. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-8624.00304.

[92] Wellman, P. H. M. *Making Minds: How Theory of Mind Develops*. Oxford University Press, Oct. 2014. Google-Books-ID: vWrDBAAAQBAJ.

[93] Wiering, M., and van Otterlo, M., Eds. *Reinforcement Learning: State-of-the-Art*, vol. 12 of *Adaptation, Learning, and Optimization*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[94] Williams, J., Fiore, S. M., and Jentsch, F. Supporting Artificial Social Intelligence With Theory of Mind. *Frontiers in Artificial Intelligence 5* (2022).

[95] Wimmer, H., and Perner, J. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition 13*, 1 (Jan. 1983), 103–128.

[96] Winfield, A. F. T. Experiments in Artificial Theory of Mind: From Safety to Story-Telling. *Frontiers in Robotics and AI 5* (2018).

[97] Wolpert, D., and Macready, W. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation 1*, 1 (Apr. 1997), 67–82. Conference Name: IEEE Transactions on Evolutionary Computation.

[98] Wu, S., and Keysar, B. The Effect of Culture on Perspective Taking. *Psychological Science 18*, 7 (July 2007), 600–606. Publisher: SAGE Publications Inc.

[99] Yang, G.-Z., Bellingham, J., Dupont, P., Fischer, P., Floridi, L., Full, R., Jacobstein, N., Kumar, V., McNutt, M., Merrifield, R., Nelson, B., Scassellati, B., Taddeo, M., Taylor, R., Veloso, M., Wang, Z., and Wood, R. The Grand Challenges of Science Robotics. *Science Robotics 3* (Jan. 2018), eaar7650.