Equipping Companions with Theory of Mind

Gregor Sieber

Introduction

Theory of Mind (ToM) is a fundamental mechanism of the human mind, allowing us to attribute mental states to other beings [Premack and Woodruff, 1978]. These mental states include beliefs, desires, intentions, and goals. Reasoning about the beliefs and desires of others is not only an important step towards building a conscious machine; several authors (e.g. [Krämer, 2008]) have also noted that it is a prerequisite for improving the communication with embodied agents.

Within the field of artificial agents and robotics, several implementations and uses of ToM have been published in the past years. ToM has mainly been used for applications in simulations and multi-agent environments, such as modeling agent-to-agent influencing [Pynadath and Marsella, 2005], self-deception [Ito et al., 2010] ,simulation based story generation using social influencing [Chang and Soo, 2008], or in game scenarios where non-player characters use ToM knowledge to improve their performance [Harbers et al., 2009, Hoogendoorn and Soumokil, 2010]

In robotics, [Scassellati, 2002] discuss two different theories of ToM and presented developments towards implementing parts of these theories for robots. [Kim and Lipson, 2009] shows how other's models of self can be learned using artificial neural networks. [Ono and Imai, 2000] demonstrate that taking into account the activation of human ToM mechanisms increases the chance of a robot's intention being understood. Also, ToM mechanisms have been used to explain the actions of an agent. Most of these implementations are concerned with a context of a limited set of goals, beliefs and desires, this set often being identical for all agents involved in the scenario.

For dialog systems geared towards long-term interaction, we envisage using ToM models in two ways: First, enriching a system with a conception of the user's mental states can improve dialog planning and enable the system to better understand the effects of its actions. Second, human ToM capabilities should be taken into account when designing the system and its interactions with a user. This means that a desired model of how its designated users should understand the system should be integrated. This enables the system to monitor whether the actions of the user - and the assumed perception of the system by the user - is in accordance with this model, by relating user input and feedback, along with derived plans, goals, and beliefs, to the model. Thus, the system will be able to react if it assumes that misconceptions of its functions, as encoded in the desired model of understanding, appear in the interaction. This will help the system avoid errors that are caused by false expectations towards its functions. Our approach suggests using two levels of ToM: the first level is concerned with

the system modeling the beliefs and goals of the user. The second level deals with what the system assumes the user's beliefs about its own beliefs are.

Background

Human Theory of Mind

Theory of Mind is present in humans in parts already in early stages of development, and fully developed around age 4. The term theory of mind was coined mainly by [Premack and Woodruff, 1978]. The two most widely accepted theories of ToM are those of [Baron-Cohen, 1995] and [Leslie, 1987] , which will be shortly described in the following paragraphs. There is also evidence of non-human theory of mind, see e.g [Call and Tomasello, 2008] for an overview.

The theory of Leslie [Leslie, 1994, 1987] was developed based on studies on the perception of causality of events in infants, compared with similar tests in adults. The studies used film sequences of colored bricks that affect each other in different ways, where some of the events violated mechanical laws. Leslie states that the world is decomposed into three distinct classes of events. The classification is based on their causal structure, these being: mechanical agency, actional agency, and attitudinal agency. Mechanical agency describes events that may be explained using rules of mechanics. Actional agency is the class of events that can be described using goals and intents of agents. The third class, attitudinal agency, covers those events that are caused by the beliefs and attitudes of an agent. For each of these classes, we are evolutionally equipped with a specific module that is able to interpret the type of event.

The Theory of Body Module handles mechanical agency, allowing humans to understand physical causality between objects. On a basic level, it is able to provide causal explanations of motion--based events such as one ball pushing away another ball it collides with. A second, more advanced part of this module is believed to operate on object--centered, three--dimensional representations generated by higher level cognitive and visual systems.

Actional and attitudinal agency -- i.e., events depending on psychological processes of agents -- are covered by the Theory of Mind Mechanism (ToMM). According to Leslie, ToMM develops in two phases, called system-1 and system-2. The first, system-1, is able to construct representations of behaviours like approach, escape, and avoidance, which result from goal-directed behaviour of agents. System-2 covers attitudinal agency. This means it represents beliefs and mental states, and their influence on agents' behaviours in relation with their goals. In contrast to system-1, system-2 allows the construction of representations where the truth values of statements are not based on observed stimuli, but on mental states. This allows for representing beliefs of others that are based on observations or knowledge that differ from our own, and forms the basis for understanding perspective and pretense.

The first of the four modules is the Intentionality Detector, which allows the interpretation of stimuli towards goals and desires. Movements of approach and avoidance can be classified. Intentionality detection only applies to stimuli with self-propelled motion, i.e. animate agents.

Second, the Eye direction detector (EDD) is used to process any stimuli that are similar to eyes and attempts to determine the direction of gaze. According to Baron-Cohen, the module serves three functions:
1) detect the presence of eye-like stimuli
2) decide whether the gaze is directed at itself or not
3) encode this perceptual state as a dyadic representation.
The shared attention mechanism (SAM) builds upon the information gained from the previous modules to create a triadic representation of the type ``person-x sees (I see person-y)". This triadic representation is created by embedding the dyadic representations of ID and EDD into each other, expressing the shared attention of the external agent and the self. SAM allows the gaze of others to be interpreted as a goal state.
Finally, the Theory of Mind Module (ToMM) provides a way of representing mental states in other agents, allowing the suspension of normal truth relations for propositions, allowing the representation of knowledge states that may contradict an agent's personal knowledge.

Note that while humans are able to make use of complex reasoning using ToM, there is evidence that they do not always do so, as they tend to simply project their own goals and beliefs onto other agents. Further background on levels of ToM and limitations on ToM use in adult humans can be found in [Keysar et al., 2003, Verbrugge and Mol, 2008, Goodie et al., 2010].

Computational models of human ToM capabilities

There have been various works on providing computational models that mimic human use of theory of mind. For example, [Ito et al., 2010] model self-deception within a decision-theoretic framework. Their work builds on an extension of subjective expected utility towards subjective expected belief utility, using decision rules for optimistic and pessimistic self-deception considering both rational belief and the desired belief. By simulating a game--theoretic scenario of action coordination, the authors demonstrate that their system is able to operationalize optimistic and pessimistic self-deception processes.

[Goodie et al., 2010] created a POMDP model of second level reasoning in competitive games based on data collected during a user study on recursive reasoning during sequential general-sum and fixed-sum two-player games. Results of the study indicate that humans are more likely to use first-level recursive reasoning than second-level reasoning in strategic general-sum games, while in more competitive and simpler fixed-sum games, the second level of reasoning is more widespread, thus attributing recursive thinking to their opponents. [Hiatt and Trafton, 2010] present an implemented cognitive model of theory of mind reasoning using the ACT-R platform. [Baker et al., 2009] present a model of human action understanding formalized as Bayesian inverse planning using Markov decision problems. Their model is evaluated using three experiments on goal inference involving agents moving in simple maze-like environments.

ToM for artificial agents

Previously, models and implementations of Theory of Mind have been used for the following applications.

- simulation of agent-to-agent influencing in decision-theoretic frameworks
- simulation of self-deception in d.-t.f
- simulation-based story generation involving social influencing
- improving performance of agents in training applications (more realistic, better simulate human reasoning, allow for explanation of behaviour and errors) using a BDI framework

[Scassellati, 2002] describes the models of Leslie and Baron--Cohen with respect to their advantages and disadvantages for implementation of a humanoid robot. While Leslie provides a very clear distinction of the perceptual world in terms of animacy, Baron--Cohen focuses more strongly on specifying the perceptual inputs for each of the modules. Based on these two theories, the author investigates the initial steps necessary for implementing a ToM in a robot. As a basis, a robot equipped with a visual, auditory, vestibular and kinesthetic system is employed. The robot is equipped with a set of basic cognitive and behavioral skills resembling the perceptual capacities of young infants: pre-attentive visual routines are able to detect color saliency, motion, and skin color. Visual attention routines model patterns of visual search and attention based on motivational influences as present in human adults. On top of that, modules for both face and eye detection, and discrimination of animate from inanimate objects are present. In the ongoing work described in the paper, the author pursued implementations of several prerequisites of a theory of mind module: gaze following, following deictic gestures, and distinction of spatio--temporal relations.

[Marsella and Pynadath, 2005] present PsychSim, a multi-agent based simulation tool that uses recursive decision-theoretic models for exploring interactions and influences between agents. As an application scenario, the authors model a school violence scenario. Agents in PsychSim each maintain their own model of the world. The *state* of an agent covers objective facts about its environment. These are features and binary relations represented as vectors with values between -1 and 1, such as the trust level between two agents or the strength of an agent. *Actions* represent what the agent is able to do. They consist of the actor, the action type, and possibly the object of the action. The agents' *goals* are represented using a reward function. Agents can have the goal of either maximize/minimize the value of a single feature, or the number of occurrences of an action. The current goals of each agent are represented using a vector of weights on the state vector representing the world. Thus, modification of the weights on different goals changes the motivation and behaviour of the agent.

Beliefs are modeled using partially observable Markov decision problems (POMDP), with limited nesting of beliefs. The belief model is composed of a recursive structure of belief vectors. Due to evidence from research on human ToM, the level of recursion can be limited to two levels, as deeper models are rarely used. This means that in the simulation, each agent has a model of: it's own view on the state of the world and itself, which may differ from its actual capabilities (e.g. an agent overestimating its own strength); and of the view of all the other agents on the state of the world

The agent's view of itself is kept separate from actual agent (i.e. may differ). Computationally, the approach is based on POMDP policy creation simplified by using mental models such as

selfishness and altruism. Reactive policies are modeled as a table of condition->action rules. While the agent uses full look-ahead for itself, the models of others are assumed to be driven by reactive behaviour. This results in more shallow reasoning and less complex models.

[Friedlander and Franklin, 2008] describes a model of theory of mind for the LIDA cognitive architecture, which is based on Global Workspace Theory. As an application, the paper describes qualitatively how findings from an experiment on Rhesus monkeys could be simulated. In the LIDA model, sensor input is passed through feature detectors, and a model of the agent's situation is constructed in the workspace. This model is enriched with associations from episodic memory and perceptual memory. Parts of the model compete for attention. The winning coalition of structures from the model is then received e.g. by procedural memory. Procedural memory contains information on possible actions, their outcomes, and activation values for these actions. Representations of other agents are created in procedural memory and may be stored in episodic memory if their activation is high enough, allowing them to be reconstructed. [Friedlander and Franklin, 2008] describes two approaches to implementing ToM in this architecture. First, the model creates similar percepts for observations of another agent to those for observations of its own actions, with the difference being their attachment to different nodes representing others or the agent's self. Second, the operations and reasoning capabilities provided by procedural memory can be applied to percepts of other agents in the way as to the agent's own. In combination, this allows reasoning about the actions of other agents.

In their application, the first-order ToM model is used to:

- query (believed) belief base of other agent
- query goal base of other agent
- predict action of other agent
- interpret action of other agent
- plan actions based on removal or addition of beliefs to other agent's believed belief/goal base (i.e. find out how to influence agent and what actions could lead there)

[Chang and Soo, 2008] state that emergent narrative requires the capability for social influence and thus social planning. They describe a model of narrative sequence (causal and problem-solving sequences) and a simulation system to reconstruct a simplified version of Shakespeare's ``Othello''. The motivations used for the agents are: greed, curiosity, jealousy, obedience. The agents are equipped with models of beliefs and goals. 2 sets of rules are used in the system: belief-revision rules update beliefs with new percepts, while motivation rules generate goals from beliefs. The agents are thus able to create ``social plans'' i.e. plans that includes actions by other agents. The implementation is based on a PDDL planner (Optop) and JADE agents. In the simulations presented by the authors, only 1 agent (``Iago'') is a social planning agents, the others are non-social. The authors' work can be seen as related to the conversation planner of Field and Ramsay (2006). The difference to frameworks based on Markov Models or decision theoretic approaches is that in the authors' system more complex plans are possible (e.g. comprising 14 steps) as opposed to the two-step look-ahead in utility maximization.

[Kim and Lipson, 2009] showed how robots that previously learned trajectories using neural networks can learn models of other robots' internal models of self and predict their future behaviour. The models were tested on simulated and real physical robots.

[Hoogendoorn and Soumokil, 2010] implemented a non-player charachter agent that is able to utilize explicit theory of mind reasoning in a BDI framework. In a first step, the agent transfers the BDI model to a default theory. In a second step, the agent reasons over the default model by calculating possible stable sets of parameters by changing the set of observations that can be manipulated in the given situation. In the final step consists in transferring observations to the plan base. The authors conducted a study comparing a simple reactive agent and a memory-based agent with 15 participants, where the ToM agent performed significantly better than the reactive agent.

Finally, there is also literature available that discuss the effect of human theory of mind on human--computer interaction scenarios. We will shortly present two studies concerned with human--robot interaction.

The first publication, [Lee et al., 2005] shows experiments about expectations towards robots depending on previous knowledge of the human user. The research question is whether people have a theory or model of how robots behave, or whether they assume the same theory as in human cooperation partners. If they do, there are likely to emerge task-specific or situation specific interaction patterns. In their experiments, participants were told the origin of a robot and asked to estimate the robot's knowledge of landmarks in two different locations. Results showed consistent with a study of the same topic about estimating human knowledge based on the same factors, and suggest that humans attribute a predictable mental model to robots from minimal information.

The second study by [Ono and Imai, 2000] shows that for HRI, ToMM needs to be taken into account on both sides. The paper suggests that successful communication depends on activating ToMM responses in the human communication partner. The authors show that such a method allows the users to understand a robot's intention even from hardly intelligible speech signals.

Implications for Companions

Resulting from an analysis of the available literature and the interaction data collected e.g. during the SERA project, we envisage the potential uses of ToM models for dialogue systems geared towards long-term interaction as described below. These potential uses should be able to contribute to the following desired features of such companions:

- be able to generate output depending on dialogue goals - related with the system's purpose - and beliefs about the world
- detect and react to topic changes induced by the user
- explain actions by being able to relate sensor input and memories to past and current belief states, and the resulting goals and actions
- be able to talk about these states, correct parts if necessary, and take these changes into account for evaluating current situation and future plans

- thus, be able to maintain multiple belief states for different points in tim, otherwise, correcting things from past interactions will make the behaviour of the system internally inconsistent with its decision mechanisms. I.e., the system needs to be able to reflect about errors it has made, while being able to use updated knowledge for deciding on the next dialog move.

As the first potential use, we believe it to be beneficial to integrate ToM into behaviour planning, as a result enabling the system to understand and create social influence processes. This may be of special benefit for applications that aim at a behaviour change, or encouragement towards specific behaviours, such as health applications.

Second, the interaction with companionable dialog systems will become more natural and efficient when *human* ToM capabilities are taken into account when designing the system and the interaction processes. This means that the design process includes the conceptualisation of a desired model of how the user should understand what the system does, based on a persona representing the potential user group of the system. Further, the system needs to be equipped with capabilities to monitor whether the actions of the user, and the thus assumed perception of the system by the user, are in accordance with this model. This can be accomplished by relating user input and feedback, along with derived plans, goals, beliefs, to the model.

Thus, this approach would use two levels of ToM: the first level is concerned with modeling the beliefs and goals of the user from the point of view of the system. The second level is concerned with what the system believes the user to believe about itself. The system then needs to be able to detect differences between the desired way of being perceived by the user, and its beliefs about how it is actually perceived by the user, and react accordingly.

Challenges:

- Often, it is an unanswered question of how we actually *want* the user to perceive a given system.
- How can the user's reactions and inputs be mapped to different belief states? What sensor data, what features of text or speech input can be used, and what memory structures and what operations on memories are required?
- What is the topology of errors that are likely to appear, and what are the best methods for handling them, taking into account research on error handling strategies?
- What approach to dialogue management should be preferred, with respect to the models it allows? As we have seen from the literature, implementations of theory of mind range from POMDPs over decision-theoretic frameworks to BDI models.

A model of theory of mind for an embodied companion system would need to link representation of intentions and beliefs to sensory input it receives from the user and its environments. There may be abstract goals that cannot be directly linked to specific sensory inputs, just like some symbols and concepts can only be related to lower--level symbols anchored in sensor input. Such goals need to be made recognizable by the agent in terms of a set of necessary conditions by decomposition into sub-goals.

Relation of Memory and ToM

Episodic Memory and Theory of Mind are thought not to be connected in humans, since ToM makes use of general knowledge encoded in semantic memory. Yet, at some point this general knowledge has been acquired from the experiences of the agent, and reasoning based on these experiences. Thus the question is whether it is possible to provide this general knowledge correctly a priori when designing an artificial system, so it fits with the world/embodiment of the system. If not, linking ToM with a mechanism that is able to acquire - and update - such rules based on episodic and autobiographic experiences of the system could provide a reasonable basis.

Representations

In order to reason using nested beliefs of the user (or other agents) these need to be represented in a way that does not conflict with or override the beliefs of the agent itself. In addition to maintaining these separate models, some of the knowledge gained from observing and modeling nested beliefs of communication partners, but also from observing their actions and interactions, may need to be added to the knowledge base of the agent. For example, the agent may find out certain properties of the user from observation or by reasoning on its memories -- e.g., that the user hardly ever gets up before 9 a.m. -- while knowledge of other properties may have been received by direct communication from the user, e.g. the user telling the agent that her favorite color is red. In this case, we argue that the agent should be able to differentiate between the different sources of knowledge, as this may influence the way the information is treated in the agent's reasoning. Different sources may indicate different levels of trust and accuracy. Knowledge from some sources may be subject to change whenever the situation requires it, while other knowledge is assumed to be constant. This is also of importance for the introspective capabilities of an agent when trying to explain behaviours and correct errors. Some sources may be identified as being the cause of an error with a higher likelihood than others, e.g. knowledge gained by information extraction methods from on--line data compared with carefully hand--crafted domain knowledge bases.

As basic categories of knowledge sources, we suggest the following categories:

- system-level basic assumptions
- initial knowledge sources
- extracted or generated by the system from interaction
- extracted or generated using tools (e.g. information extraction, www search)
- provided by other agents or the users

When storing shared experiences of the user and the agent in memory, it is crucial to not only store the factual state of the situation, but to firstly separate the observable facts from the agent's interpretation of the situation, and secondly to store the likely interpretations of the user, and whether the situation was positive or negative for the user, contributed to a goal, was potentially face-threatening, etc.

According to [Nickerson, 1999] , a default measure humans apply for assessing what another person knows is to assume that they have the same knowledge as oneself. While this may lead

to a number of errors, it can be seen as a valid mechanism when nothing else is known, especially given the amount of knowledge humans share that is based on the experience of living in human society. This can be seen as a baseline function when implementing ToM in a companion agent.

The first requirement for implementations of ToM is a detailed model of an agent's own knowledge. The contents of its memory need to be annotated with information on where the content comes from, whether it is a) from a machine viewpoint or human knowledge and b) specific to this system or general knowledge. knowledge sources or functions can be e.g. be done using names to classify a user as belonging to a certain class, the probability of this knowledge being common ground increases.

Potential methods of representing ToM knowledge in RDF

The main issue to tackle when representing ToM-related knowledge in RDF is to track the provenance of statements, i.e. where they came from, and what role they play. Note that generally, in an RDF triple store, each triple comes on its own and there is no built--in mechanism that allows the association of meta-information. This also requires additional mechanisms e.g. for temporal grouping of statements, e.g. for representing everything associated with one given action.

The technical possibilities can generally be divided into the following categories:

1. Namespaces
2. Reification
3. Named graphs
4. Extensions of the triple model to a quadruple model

Namespaces are a basic mechanism of identifying where the semantics of a statement is defined, and where its origin is. This would be an option to identify e.g. purely system-specific knowledge. For representing knowledge that is potentially inconsistent, namespaces do not seem like a good option.

As the standard mechanism for representing meta-information in RDF, reification is used. Thus, statements about statements are created. When used heavily, this creates significant overhead and complicates RDF retrieval/querying and inference.

As an alternative to tracking provenance or trust of statements using reification, there have been extensions of the triple model to quadruple models. Named graphs can be seen as one such option, but other implementations are a topic of current research to provide more flexibility. The Sesame RDF architecture supports named graphs by providing so-called context value for each statement (i.e. quadruple model). This context value could be used to efficiently represent the different sources of knowledge, or different classes of reliability. For retrieval of statements, the context can be limited to retrieve only statements with a certain reliability or source.

References

[Baker et al., 2009] Baker, C., Saxe, R., and Tenenbaum, J. (2009). Action understanding as inverse planning. Cognition, 113(3):329-349.

[Baron-Cohen, 1995] Baron-Cohen, S. (1995). Mindblindness: An essay on autism and theory of mind. The MIT Press.

[Call and Tomasello, 2008] Call, J. and Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. Trends in Cognitive Sciences, 12(5):187-192.

[Chang and Soo, 2008] Chang, H. and Soo, V. (2008). Simulation-based story generation with a theory of mind. In Proceedings of the Fourth Artificial Intelligence and Interactive Digital Entertainment International Conference (AIIDE 2008), pages 16-21.

[Friedlander and Franklin, 2008] Friedlander, D. and Franklin, S. (2008). LIDA and a Theory of Mind. In Proceeding of the 2008 conference on Artificial General Intelligence 2008: Proceedings of the First AGI Conference, pages 137-148. IOS Press.

[Goodie et al., 2010] Goodie, A., Doshi, P., and Young, D. (2010). Levels of theory-of-mind reasoning in competitive games. Journal of Behavioral Decision Making.

[Harbers et al., 2009] Harbers, M., Van den Bosch, K., and Meyer, J. (2009). Enhancing training by using agents with a theory of mind. Proceedings of EduMas, pages 23-30.

[Hiatt and Trafton, 2010] Hiatt, L. M. and Trafton, J. G. (2010). A cognitive model of theory of mind. In Salvucci, D. D. and Gunzelmann, G., editors, Proceedings of the 10th International Conference on Cognitive Modeling, Philadelphia, PA. Drexel University.

[Hoogendoorn and Soumokil, 2010] Hoogendoorn, M. and Soumokil, J. (2010). Evaluation of virtual agents utilizing theory of mind in a real time action game. In Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems, volume 1, pages 59-66. International Foundation for Autonomous Agents and Multiagent Systems.

[Ito et al., 2010] Ito, J., Pynadath, D., and Marsella, S. (2010). Modeling self-deception within a decision-theoretic framework. Autonomous Agents and Multi-Agent Systems, 20(1):3-13.

[Keysar et al., 2003] Keysar, B., Lin, S., and Barr, D. J. (2003). Limits on theory of mind use in adults. Cognition, 89(1):25-41.

[Kim and Lipson, 2009] Kim, K. and Lipson, H. (2009). Towards a simple robotic theory of mind. In Proceedings of the 9th Workshop on Performance Metrics for Intelligent Systems, pages 131-138. ACM.

[Krämer, 2008] Krämer, N. (2008). Theory of mind as a theoretical prerequisite to model communication with virtual humans. In Modeling Communication with Robots and Virtual Humans, volume 4930 of Lecture Notes in Computer Science, pages 222-240. Springer Berlin / Heidelberg.

[Lee et al.,] Lee, S.-l., Lau, I. Y.-m., Kiesler, S., and Chiu, C.-Y. (2005) Human mental models of humanoid robots. Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA 2005). April 18-22, Barcelona, Spain, pages 2767 – 2772.

[Leslie, 1987] Leslie, A. (1987). Pretense and representation: The origins of theory of mind.". Psychological review, 94(4):412-426.

[Leslie, 1994] Leslie, A. (1994). Pretending and believing: issues in the theory of ToMM* 1. Cognition, 50(1-3):211-238.

[Marsella and Pynadath, 2005] Marsella, S. and Pynadath, D. V. (2005). Modeling inuence. and theory of mind. In Arti_cial Intelligence and the Simulation of Behavior.

[Nickerson, 1999] Nickerson, R. S. (1999). How We Know-and Sometimes Misjudge-What Others Know: Imputing One's Own Knowledge to Others. Psychological Bulletin, 125(6):737-759.

[Ono and Imai, 2000] Ono, T. and Imai, M. (2000). Reading a robot's mind: A model of utterance understanding based on the theory of mind mechanism. In Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence, pages 142-148. AAAI Press.

[Premack and Woodruff, 1978] Premack, D. and Woodru_, G. (1978). Does the chimpanzee have a theory of mind? Behavioral and Brain sciences, 1(04):515-526.

[Pynadath and Marsella, 2005] Pynadath, D. V. and Marsella, S. C. (2005). Psychsim: modeling theory of mind with decision-theoretic agents. In Proceedings of the 19th international joint conference on Artificial intelligence, pages 1181{1186, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Scassellati, 2002] Scassellati, B. (2002). Theory of mind for a humanoid robot. Autonomous Robots, 12(1):13-24.

[Verbrugge and Mol, 2008] Verbrugge, R. and Mol, L. (2008). Learning to apply theory of mind. J. of Logic, Lang. and Inf., 17:489-511.