DiVA
http://www.diva-portal.org

This is the published version of a paper published in *Paladyn - Journal of Behavioral Robotics*.

Citation for the original published paper (version of record):

Hellström, T., Bensch, S. (2018)
Understandable Robots: What, Why, and How
Paladyn - Journal of Behavioral Robotics, 9(1): 110-123
https://doi.org/10.1515/pjbr-2018-0009

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:
http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-150156

Research Article **Open Access**

Thomas Hellström* and Suna Bensch

Understandable robots

- What, Why, and How

https://doi.org/10.1515/pjbr-2018-0009
Received December 14, 2017; accepted May 17, 2018

Abstract: As robots become more and more capable and autonomous, there is an increasing need for humans to understand what the robots do and think. In this paper, we investigate what such understanding means and includes, and how robots can be designed to support understanding. After an in-depth survey of related earlier work, we discuss examples showing that understanding includes not only the intentions of the robot, but also desires, knowledge, beliefs, emotions, perceptions, capabilities, and limitations of the robot. The term *understanding* is formally defined, and the term *communicative actions* is defined to denote the various ways in which a robot may support a human's understanding of the robot. A novel model of interaction for understanding is presented. The model describes how both human and robot may utilize a first or higher-order theory of mind to understand each other and perform communicative actions in order to support the other's understanding. It also describes simpler cases in which the robot performs static communicative actions in order to support the human's understanding of the robot. In general, communicative actions performed by the robot aim at reducing the mismatch between the mind of the robot, and the robot's inferred model of the human's model of the mind of the robot. Based on the proposed model, a set of questions are formulated, to serve as support when developing and implementing the model in real interacting robots.

Keywords: human-robot interaction, communication, predictable, explainable

***Corresponding Author: Thomas Hellström:** Department of Computing Science, Umeå University, Sweden,
E-mail: thomas.hellstrom@umu.se
Suna Bensch: Department of Computing Science, Umeå University, Sweden, E-mail: suna.bensch@umu.se

1 Introduction

Several research efforts within HRI focus on how to make robots understand human actions and thoughts. In particular, techniques for recognition of human activities, intentions, and emotional states have been developed based on, for example, analysis of human motion, gestures, facial expressions, and verbal utterances. This line of research is as such highly important and far from completed. However, as robots become more and more competent and autonomous, there is an increasing need to study also how humans understand robots. This need has been acknowledged also for Artificial Intelligence (AI) in general, recently under the notion of "Explainable AI" [1, 2]. In particular social robots, that work closely with humans, should be designed such that the humans understand how the robots think and act [3]. Failure to address this perspective may negatively affect interaction quality [4] between a robot and its users, and degrade user experience, efficiency, and safety. A robot that acts without communicating its intentions to involved humans may create anxiety [5], just as a human behaving the same way does. In a cooperative task, the overall efficiency may be negatively affected if the human cannot correctly anticipate the actions of the robot [6, 7]. In non-cooperative tasks, the robot may fail to complete a commanded task if it does not inform the human about its internal state, e.g. the need to charge the batteries. Safety may be negatively affected in several respects. The risk for physical collisions increases if the human is unaware of the limited view from the perspective of the robot, and the possibilities for the human to correct or stop unsafe robot actions are limited if the human cannot understand and predict the motion of the robot. One important and active application area is autonomous cars [8]. Experiments have shown that pedestrians feel unsafe and experience discomfort and confusion when they encounter autonomous vehicles [9], and ways to communicate the intention of autonomous vehicles are being developed by the automotive industry, for example in the AVIP project (https://www.viktoria.se/projects/avip-automated-vehicle-interaction-principles). Likewise, com-

Open Access. © 2018 Thomas Hellström and Suna Bensch, published by De Gruyter. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 License.
Brought to you by | Umea University Library
Authenticated
Download Date | 10/26/18 4:01 PM

municating the reliability of a system helps users to calibrate their trust of the system [10].

While understandability often is the goal of Human-Robot Interaction (HRI) research, an analysis of what the concept really means and how it can be formalized is to the authors' knowledge missing. This paper aims at filling this gap, thereby providing a foundation for continued research on the topic. After an in-depth survey of related earlier work in Section 2, Section 3 analyses what "understanding a robot" means. In Section 4, a novel model of interaction for understanding is presented. Based on this model, general guidelines for design of interaction for understanding are formulated in Section 5. The paper is finalized in Section 6 by a summary of findings, and conclusions regarding future work in the area.

2 Related earlier work

The importance of understandable robots has been acknowledged by the HRI community for a long time, for example in organized workshops on *explainable robots* [11]. A variety of different terms have been used. *Readability* [12], *anticipation* [7], *legibility* [13, 14], and *predictability* [15] usually refer to how humans should be able to predict a robot's future behavior, in particular physical motion. Dragan et al. [16], make a distinction between legibility and predictability, with the former being connected to understanding of goals and the latter to understanding actions. In [17], the author uses the term *intelligibility* specifically for humans' understanding of robot emotions. The authors in [18] use *intent communication* as a more general term to denote how robots communicate goals (object or aim) and also the reason for pursuing these goals. Mirnig and Tscheligi [19] argue that, in particular social robots must be able to communicate its internal system status to interacting humans, through "active feedback", i.e. output deliberately created by the robot for the purpose of understanding. Knifka [20] discusses work by, among others, Breazeal [21] and Dautenhahn [22], and uses the term *understanding* to cover not only physical but also social interaction (Breazeal [21] also uses the word *readability* for this). In [23], the word *transparency* is used to describe roughly the same thing, and Lyons [24] discuss how such transparency can be achieved not only through design of the human-robot interface but also through training of users on the robot system. In this paper, we use the terms *understanding* and *understandability* with a formal definition given in Section 3.

In the robotics community, related earlier work is done in the area of *intention recognition*, for example through a series of international workshops at HRI, HAI, and RO-MAN. Some research address how humans recognize intentions of robots, and how the robots may support that process. However, most research in intention recognition describe specific techniques by which robots can recognize intentions of humans. Furthermore, intention is, as we exemplify further on, only one part of general understanding.

There is a tight connection between understanding and communication [25], and our proposed model of interaction for understanding will build on models for general communication.

The remaining survey of related earlier work is divided into four topics: communication for understanding, humans understanding humans, humans understanding robots, and robots understanding humans. This division of work is chosen although humans' and robots' understanding of each other is often highly intertwined. For example, one important part of a robot's understanding of a human, may be to understand the human's understanding of the robot. This interdependence is explicit in our proposed model, but for now we review earlier work in each area separately.

2.1 Communication for understanding

Our proposed model of interaction for understanding will build on existing theories on communication, modified to fit the specific case of communication that supports understanding in HRI. A starting point will be the model suggested by Shannon in 1948 [26]. According to this model, communication can be conceptualized as shown in Fig. 1. An information source produces a message to be communicated. A message transmitter encodes (translates) the message into signals. A channel is used to transfer the signals, possibly corrupted by noise, to a receiver, which decodes (translates back) the message and delivers it to the destination. This model has been highly influential but also criticized of being inappropriate for social sciences, and in particular, to model interpersonal communication. For example, Chandler [27] pointed out that the model employs a "postal metaphor" by describing communication as sending a physical package of information to a receiver. Hence, communication would be essentially one-way and linear, with an active sender and a passive receiver. This view is not considered as appropriate to describe interpersonal communication that extensively adapt to responses and cues from the receiver. To some extent, this can be

Brought to you by | Umea University Library
Authenticated
Download Date | 10/26/18 4:01 PM

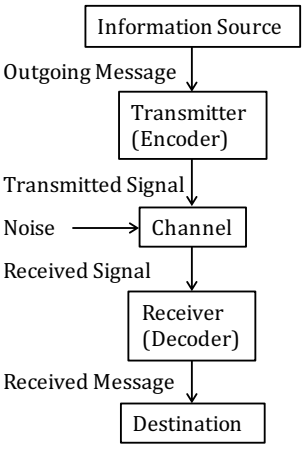

Figure 1: Schematic diagram of a general communication system, as described by Shannon in [26].

taken into account by adding a feedback loop in the model, as also suggested in a modified model by Schramm [28]. Chandler further criticizes Shannon's model for not taking *meaning* into account. In particular it does not reflect how the decoding phase in interpersonal communication depends heavily on context and the receiver's social and cultural environment. Instead, decoding is assumed to fully recover the transmitted message (in the absence of noise). However, in a real-world setting, effective communication can only be achieved if the sender takes into account how the receiver decodes and interprets the received message.

2.2 Humans understanding humans

Humans daily use mindreading in order to estimate the mental states and actions of others by means of observing their behaviors, and this ability is regarded as most essential for the success of the human species [29]. Baron-Cohen describes mindreading as a necessary survival strategy from an evolutionary perspective [30, p. 25]. The concept *theory of mind* (ToM) (also known as *mindreading* or *mentalizing*) is used to denote the ability to attribute a mind with mental states (beliefs, intents, desires, pretending, knowledge, etc.) to oneself and others, and to understand that others have beliefs, desires, intentions, and perspectives that are different from one's own [31]. Michlmayr identifies three important functions of a ToM [32]: 1) To comprehend and explain the behavior of other people. Without it, we may get confused and overwhelmed by the complexity of the world. 2) To predict other's behavior, which can be seen as a general requirement to deal with other people [33, p. 57]. 3) To manipulate and influence others by controlling the information available to them. For this to be possible, other's goals, desires and beliefs have to be perceived. Also, estimating someone else's estimation of your mind is beneficial. For example, what a person expects you to do may very well affect her behavior, and being able to predict that may give you certain advantages. This mechanism is called *second-order theory of mind* and can be further extended to higher orders [34]. The term *zeroth-order theory of mind* is sometimes used to denote reasoning based on only the agent's own mind, without taking into account others' mind. In the following, the expression *theory of mind* (ToM), denotes first-order theory of mind unless otherwise stated.

With humans, the development of a ToM starts already at infancy and continues through to the adolescent years [33]. Whether other animals than humans have or can acquire a ToM is an open question [31], even if some evidence indicate that chimpanzees [35] and even birds like ravens [36] might possess this ability to "put themselves into others' shoes".

There are two major views for how a ToM works in a human [37]. According to *simulation theory* [38], we simulate another agent's actions and stimuli using our own processing mechanisms, and can thereby predict the agent's behavior and mental state. Neurophysiological support for this theory has been found in the brain of a certain kind of monkeys [39]. So-called mirror neurons are activated both when the animal performs an action and when it perceives another monkey performing the same action. Mirror neurons are therefore supposed to be a major part of the simulation mechanism. In humans, individual mirror neurons have not been identified, but regions of neurons have been shown to behave in the same way [39]. The other major view is referred to as the *theory theory*. This theory builds on the assumption that we are equipped with a "folk psychological" theory consisting of a set of laws and rules that connect mental states with sensory stimuli and behavioral responses. These rules can be used to understand others' mental states and behaviors [40, p. 207]. Rules can be in the form of causal laws, such as "A person denied food for any length will feel hunger" [33, p. 53]. Rules can also be in the form of general principles such as the "Law of the practical syllogism" [38]: "If S desires a certain outcome G and S believes that by performing a certain action A she will obtain G, then ceteris paribus S will decide to perform A". Some researchers hold that a folk psychological theory is learned as a child grows up [33], while others hold a nativistic attitude [41].

Approaches developed for robots' understanding of humans' intentions have also been suggested as explanation of humans' recognition of the behavior of others. Inverse planning relies on a "principle of rationality", i.e. the

Brought to you by | Umea University Library
Authenticated
Download Date | 10/26/18 4:01 PM

assumption that all actions aim at efficiently reaching a set goal [29], which makes it possible to infer an agent's goal by observing its actions. Research suggests that humans use inverse planning to infer goals and intentions from observed actions. For example, psychological studies of preverbal infants indicate that they infer plans from observed sequences of human actions [42] (see [29] and [43] for comprehensive overviews).

2.3 Humans understanding robots

A large body of research deals with how robots can be made understandable by equipping them with static functions that provide information about the robots' state and behavior. This kind of robots do not incorporate any ToM of the interacting human, and information is provided without considering the human's varying need of information for understanding. A few examples of this approach are given below (more examples are given in Section 3). The authors in [44] describe a prototype of a robotic fork-lift equipped with a projector to visualize internal states and intents on the floor. A similar approach is taken in [45] with an assembly robot. The robot projects assembly information and planned robot movements onto the work table in order to improve collaboration with a human operator. In [46], an intent communication system for autonomous vehicles is presented and evaluated. A combination of strobe lights, a LED word display, and speakers are used to communicate whether the car would like a pedestrian to cross or stop. In [47], three types of signals were used to inform an interacting human about the robot's intention to sweep the area under an obstacle or the need for user's help to remove the obstacle.

To extend this kind of static provision of information, design of understandable robots may exploit human anthropomorphism. It is well known that humans have a tendency to attribute minds not only to other humans but also to non-human animals, dead objects, and even natural phenomena such as thunder in our attempts to predict future states of the world. Dennett [48] denotes this as taking an "intentional stance" toward things. Research using neuroimaging indicate that the tendency to attribute a mind also holds for interaction with robots, and possibly depends on embodiment and human likeness [49]. Recent research [50] shows that people spontaneously attribute mental states to a robot by taking the robot's visual perspective, which is one of the most significant ToM precursors. Experiments reported in [51] indicate that people form mental models of a robot's factual knowledge based on their own knowledge, and on information about the robot's origin and language. Such mechanisms may be used by robots to communicate internal states, and affect human emotions and views of robots. In [47], a robot was designed to move back and forth to show the user that it wanted an obstacle to be removed. The results showed that this "emotional" behavior encouraged most users to help the robot. Experiments reported in [52] show how the reactions of robots when being touched can be used to create impressions of familiarity and intentionality with the interacting human. Experiments reported in [92] show that a robot may use eye movements or body poses (such as leaning back or craning its neck) to communicate its opinion on personal space around the robot. The timing of behaviors carries information that interacting humans may interpret in various ways. In [53], a robot arm was controlled to move a cup over a table for a handover configuration. Speed, change of speed, and pauses were varied and were shown to have a clear influence over how observers described the robot's behavior. For example, a slow robot was more often described as careful, cautious, or deliberate than a fast-moving robot. In [54], Ono and Imai show how a robot that gives the illusion of having a "mind", improves an interacting human's ability to interpret and act upon unclear verbal commands.

The authors of [55] emphasize the importance of having a good mental model of the robot's decision mechanism (or the *robot's objective function* as the authors denote it) in order to predict a robot's future behavior. To support this, they present an approach in which the robot models how a human infers the robot's objectives from observed behavior, and then chooses the most informative behavior to communicate the objective function to the human. This is one (rare) example of how a second-order ToM is implemented in a robot.

2.4 Robots understanding humans

How computers can analyze and understand intentions and actions of humans has been intensively studied in several areas, with varying foci and approaches. One large direction of research is *plan recognition* [56], which deals with the problem of how observed sequences of actions can be seen as (part of) a plan to reach a goal, and is typically approached with techniques for graph-covering problem [56], probabilistic inference [57], parsing [58], or Hidden Markov Models (HMMs) [59, 94]. The research community has, using the acronym PAIR (Plan, Activity, and Intent Recognition), for several years organized related workshops (see e.g. http://www.planrec.org/PAIR/Resources.html). The sub-

Brought to you by | Umea University Library
Authenticated
Download Date | 10/26/18 4:01 PM

areas activity recognition and behavior recognition deal with how observed sequences of noisy sensor data can be associated with specific actions, and are tightly connected to Learning From Demonstration [60, 61], Programming By Demonstration [62], and Imitation Learning [63, 64]. Research on intention recognition is also conducted in several other research areas. For example, intention recognition based on human utterances is done as part of Natural Language Processing [65, 66]. Intention may also be inferred from body language, gaze [67], and facial expressions.

Attempts have been made to implement a ToM in a robot to predict human mental states and actions. Scassellati was one of the first to suggest and implement ToM in robotics. In [68], he combined and implemented ToM models by Baron-Cohen [30] and Leslie [69] to construct a robot with shared visual attention. Bennighoff et al. [70] presented (non-conclusive) experimental results indicating that robots equipped with a ToM are viewed as more sympathetic by interacting humans. The research reported in [71, 72, 96], investigate how a robot can form a representation of the world from the point of view of a human by reasoning about what the human can perceive and not. In [73] Kim and Lipson describe a robot that models other's self with an artificial neural network and an evolutionary learning mechanism. Reportedly, the robot manages to successfully recover other's self-models. In [74], Devin and Alami describe a robot that models a human's mental states in order to perform joint actions with the human. The robot is able to adapt to the decisions of the human, and also informs the human without giving unnecessary information that the human can observe or infer by herself. Hiatt et al. [75] describe a robot that analyses unexpected actions by a human through simulation analysis of several hypothetical cognitive models of the human. The robot uses a ToM to model the human's knowledge and beliefs about the world, and the reported experimental results show that the robot is viewed as a more natural and intelligent teammate than with alternative approaches.

Apart from rather simplistic experiments, like the ones reviewed above, it is striking how "even the most advanced, lifelike robots cannot reason about the beliefs, desires and intentions of other agents" [29].

3 What does it mean to understand a robot?

We take a pragmatic approach to the meaning of the word "understanding" and refer to it as "… a psychological process related to an abstract or physical object, such as a person, situation, or message whereby one is able to think about it and use concepts to deal adequately with that object" [76]. More specifically, we focus on the process that enables a human to successfully interact with a robot. While the discussion and presented models apply also to more abstract levels of understanding, such as social context [20–22] and purpose [24], our examples are mostly at a lower level of abstraction.

One important aspect of understanding concerns goal-directed actions and intentions of a robot [77]. For interaction to be natural, efficient, and safe, the human must often understand what the robot does, and why it acts the way it does. A few concrete examples are:

1. A service robot that decides to take out the garbage should (in some cases) inform its owner about the planned action.
2. A mobile robot that needs a human to move aside to get through should explain both need and reason to the human [78].
3. In a pick-and-place scenario involving a robot and a collaborating human, the robot arm should move towards objects and locations in a manner predictable by the human [79].
4. An autonomous car detecting a pedestrian attempting to cross the highway should communicate whether it will cooperate by slowing down or not [8, 46].

However, understanding of a robot is not limited to physical actions and intentions, but also includes entities such as desires, knowledge and beliefs, emotions, perceptions, capabilities, and limitations of the robot [80], and also task uncertainty [81], and task progress [82, 93]. A few concrete examples of how such understanding is relevant and important are:

5. A robot that expresses emotions, for example frustration over a task [17, 47], or general needs [83], can get help in a natural way from an interacting human.
6. A service robot should inform its user about its battery status if it expects to run out of power during a planned task.
7. An autonomous car should inform the passengers about changed route plan due to an updated weather forecast.
8. A robot that acts on verbal commands from the user may express uncertainty regarding the meaning of a command, by waiting before acting, by moving more slowly, or through gestures expressing hesitation [84].

Hence, understanding of a robot may relate to both deliberate physical actions, and a large number of non-physical

Brought to you by | Umea University Library
Authenticated
Download Date | 10/26/18 4:01 PM

entities, such as the previously given examples of desires, knowledge and beliefs, emotions, perceptions, capabilities, and limitations of the robot, task uncertainty, and task progress. However, physical actions are tightly connected to non-physical entities such as intentions and goals, and we will, somewhat loosely, refer to all such entities collectively as the *state-of-mind* (SoM) of the robot. An alternative would be to simply use the word "mind", but we prefer "state-of-mind" and "SoM" to avoid misleading anthropomorphism.

With reference to the previously quoted definition of the word understanding, we introduce the following definition:

Definition 1 An agent's understanding of another agent is the extent to which the first agent has knowledge about the other agent's SoM in order to successfully interact with it.

Hence, we say that a human understands a robot if she has sufficient knowledge of the robot's SoM in order to successfully interact with it. Likewise, we say that a robot understands a human if it has sufficient knowledge of the human's SoM in order to successfully interact with it.

4 Modeling interaction for understanding

In many cases, a human may understand what a robot is doing by simply observing it moving and acting in order to reach internal goals. If these goals are related to commands by the interacting human, understanding becomes even easier. However, as robots become more and more autonomous and complex, they will also become increasingly harder to understand. A robot can support understanding by performing communicative actions that increase the interacting human's knowledge of the robot's SoM. The term communicative actions is used in [85, 86] to refer to "... behaviors that implicitly communicate information ...". We introduce a more specific definition:

Definition 2 A communicative action is an action performed by an agent, with the intention of increasing another agent's knowledge of the first agent's SoM.

Hence, a communicative action is performed by an agent to increase another agent's understanding of the first agent. In the remainder of this section we present a model of HRI specifically describing interaction for generation, communication, and interpretation of communicative actions.

It is sometimes sufficient for a robot to generate static communicative actions, such as reported in [24, 44–47, 82]. With modest requirements, the eight examples in Section 3 would work with static actions, but for higher interaction quality, communicative actions should be designed to fit the current perspective and needs of the human. For this, the robot benefits from inferring a model of the human's mind by utilizing a first-order ToM. For example, in Example 7, the autonomous car should not inform the human about the changed route plan more than once. To manage this, the car needs to estimate the human's current knowledge, i.e. it needs a ToM of the human. In Example 1, the robot may utilize a ToM to determine whether the human should be informed or is too busy to be disturbed.

Other cases require the robot to be equipped with a second-order ToM, such that the robot assumes not only that the human has a mind, but also that she has a ToM of the robot. This enables design of communicative actions that provide specific missing information to the human. One example is the already reviewed work in [55]. Another example would be Example 4, with the extra requirement that the autonomous car should infer not only the pedestrian's intention to cross the highway, but also the pedestrian's belief regarding the car's intention to brake or not. If the car's intention does not match the pedestrian's belief, the car should perform communicative actions in order to change the pedestrian's belief, or adapt its own behavior. For example, if the car initially plans to continue without braking in order to avoid collision with another car approaching from behind, and the pedestrian does not seem to pay attention to this fact, the car may reevaluate the risks and brake in order to avoid an imminent collision with the pedestrian. To accurately handle this kind of bidirectional interaction, the proposed model equips both human and robot with a mind that includes a model of the other's mind. Hence, the robot's mind contains a model of the human's mind, part of which is a model of the robot's mind (this does not necessarily lead to infinite recursion since a model at a certain level may be defined as not containing any further models).

The basic driving force to generate communicative actions is the mismatch between the robot's mind and its model of the human's model of the robot's mind. Communicative actions are generated with the goal of reducing this mismatch.

The connection between understanding and communication is well expressed in the following definition by Anderson [25]: "Communication is the process by which we understand others and in turn endeavor to be understood by them". Our model builds on and extends Shannon's model of general communication [26], earlier de-

Brought to you by | Umea University Library
Authenticated
Download Date | 10/26/18 4:01 PM

scribed in Section 2, and describes how a human and a robot act in order to support mutual understanding by generating, communicating, and interpreting communicative actions that support the human's understanding of the robot. Since this, as exemplified above, sometimes requires a second-order ToM, also the robot's understanding of the human has to be involved. The model is therefore based on two instances of Shannon's model, thereby duplicating the mechanism for encoding, transferring, and decoding such that robot and human are modeled as simultaneous senders and receivers. This design step also addresses Chandler's previously mentioned "postal metaphor" criticism [27]. The proposed model is illustrated in Fig. 2. The robot's SoM M_R contains m_H, a model of the human's SoM M_H. In a symmetric fashion, M_H contains m_R, a model of M_R. By Definition 1, human understanding of the robot relates to the mismatch between M_R and m_R. We denote this mismatch $|m_R - M_R|$. The notation should not be interpreted mathematically, but rather symbolically as a measure of the extent to which relevant parts of M_R and m_R differ. For full understanding, M_R and m_R do not necessarily have to be identical, but there should be no relevant mismatch. What is relevant and not is clearly application dependent.

Human understanding of the robot is established and supported by sequential execution of the three modules I_R, N_R, and G_R:

I_R The robot infers m_H by using M_R, communicative actions A_H generated by the human, and general interaction Ix between human and robot.

N_R The robot compares its mind M_R with its estimation of m_R, the human's model of M_R (this estimation is part of m_H). If the estimation of $|m_R - M_R|$ exceeds a set threshold, the robot identifies which information the human needs in order to reduce $|m_R - M_R|$.

G_R The robot selects, generates, and executes appropriate communicative actions A_R aiming at communicating the needed information.

The interacting human's cognitive process is modeled symmetrically in the three modules I_H, N_H, and G_H:

I_H The human infers m_R in module I_H by using M_H, communicative actions A_R generated by the robot, and general interaction Ix between human and robot.

N_H The human compares its mind M_H with its estimation of m_H, the robot's model of M_H (this estimation is part of m_R). If the estimation of $|m_H - M_H|$ exceeds a set threshold, the human identifies which information the robot needs in order to reduce $|m_H - M_H|$.

G_H The human selects, generates, and executes appropriate communicative actions A_H aiming at communicating the needed information.

While a robot can be designed to work according to proposed model, there is of course no guarantee that an interacting human would do the same. However, the robot part of the model is applicable even if the human is not interested in being understandable as suggested by the model.

As mentioned in modules I_R and I_H above, inference of m_H and m_R sometimes is done not only from communicative actions A_R and A_H, but also from general interaction Ix between human and robot. For example, human intention can sometimes be inferred by observing physical motion of the human, not specifically performed to support understanding (this is the typical case in intention recognition, behavior recognition, and activity recognition). Human interaction often involves complex mixtures of communication using both Ix and communicative actions A_H, and may also depend on corresponding interaction generated by the robot [87]. Such complex interaction modes are not explicitly described by the presented model.

Inference of m_H and m_R may also use information in M_R and M_H respectively. For example, the autonomous car in Example 7 could remember (store in M_R) that it has informed the human about the updated route plan. Using this stored information, the car may infer that the human holds a model m_R of the robot's SoM, with an updated route plan. Since there will be no critical mismatch regarding knowledge of the route plan, the car will not repeat the same communicative action over and over.

As illustrated in Fig. 2, generation of communicative actions (in G_R) mainly depends on the identified information to be communicated (N_R), but also on the information in M_R. For example, using gestures as modality for a communicative action only works if the receiver is looking at you, and speech may be a bad choice in a noisy environment. Hence, to determine appropriate modality, a robot needs to access both its model of the human's mind, and its local perception. Both are available in M_R.

The choice of appropriate communicative actions gets even more complex if the sensing and acting capabilities of human and robot are not symmetric [88]. One example is a humanoid robot with the camera mounted on its belly and not inside its eyes. A human might then try to communicate to the robot by showing gestures or signs to its eyes (i.e. similar to the mistake we sometimes make in video conferences). This problem can in some cases be dealt with by modifying the physical design of the robot. In the other direction, the robot may overcome differences

Brought to you by | Umea University Library
Authenticated
Download Date | 10/26/18 4:01 PM

in sensing and acting capabilities by customizing its communicative actions. For example, to communicate visual attention, the robot may turn its head, even if the camera is mounted on the belly. In other cases, such customization may involve learning to adapt to the human's needs and capabilities [3, p. 157].

Our proposed model benefits from the conceptual separation of message and signal suggested in Shannon's model (Fig. 1). In the same way as a message is separate from the transmitted signal, the information to be communicated is separate from the chosen communicative action. For example, just like a sequence of words can be transmitted with Morse signals on a copper wire, or with an email, so can a robot's emotional state be communicated with a body pose, or with a spoken utterance. Furthermore, the proposed model acknowledges previously expressed criticism of Shannon's model as not taking context into account when decoding messages [27]. In our model, the corresponding operation in the robot is the inference of m_H, which depends not only on the communicative action A_H, but also on general interaction Ix and the robot's SoM M_R. Hence, a given communicative action may very well be interpreted differently by the receiver, depending on additional interaction and previous perception. Just like in Shannon's model, the interpretation may be further corrupted by communication noise.

4.1 Examples of applying the model

The proposed model may be applied to the extended example with the autonomous car described in Section 4. In this example, the car's pedestrian protection system detects a pedestrian approaching the highway. Based on the traffic situation the system decides not to slow down. Generation, communication, and interpretation of communicative actions take place as follows:

I_R The car infers that the pedestrian believes that the car intends to slow down, since the pedestrian is entering the road.

N_R The car concludes that there is a serious mismatch between the car's intention and the inferred belief of the pedestrian. Communicating the car's intention to the pedestrian is chosen as a means to reduce the mismatch.

G_R The car honks and flashes the headlights as means to communicate the intention.

The pedestrian's cognitive processes are modeled as follows. The pedestrian decides to cross the road, enters it, and performs the following steps:

I_H The pedestrian interprets the honking and flashing headlights as signals indicating that the car does not intend to slow down, but rather expects the pedestrian not to proceed crossing the road. This is also the new decision made by the pedestrian (made outside of the model).

N_H The pedestrian estimates that there is no serious mismatch between m_H (the car's belief that the pedestrian will not cross) and M_H (the fact that the pedestrian does not intend to cross). Hence, there is no need to communicate any information to the car.

G_H The pedestrian performs no communicative actions. It should be noted that the pedestrian stopping may be seen as a communicative action that, in the next iteration of the model, is interpreted by the car as an intention not to cross the road.

While this example shows a complex case with second-order ToM on both the robot's and the human's side, the model also applies to simpler modes of interaction. In Example 6, the robot may be designed to inform its user about battery status (the robot's communicative action) every time it expects to run out of power, i.e. without inferring the human's state-of-mind or analyzing any communicative actions from the human. The robot would simply always assume that the human does not know about the battery status of the robot (i.e. it would assume a certain fixed mismatch between M_R and m_R). This corresponds to the robot utilizing a zeroth-order ToM of the human.

5 Designing interaction for understanding

As described in Section 4, a robot may perform communicative actions in order to be understood. Our proposed model describes in general terms how such actions are generated, communicated, and interpreted, but an actual implementation requires application specific realizations of the modules I_R, N_R, and G_R in the model. Such realizations can be guided by answering the following questions and sub questions:

Q1 **(N_R) What information (if any) should be communicated to the human?** The first thing to clarify is what understanding we want to achieve, and what the mismatch $|m_R - M_R|$ should represent and include. More specifically we need to answer:
a) How should the mismatch $|m_R - M_R|$ be estimated? Note that m_R is out of reach for the robot, and the mismatch has to be estimated from information in m_H.

Brought to you by | Umea University Library
Authenticated
Download Date | 10/26/18 4:01 PM

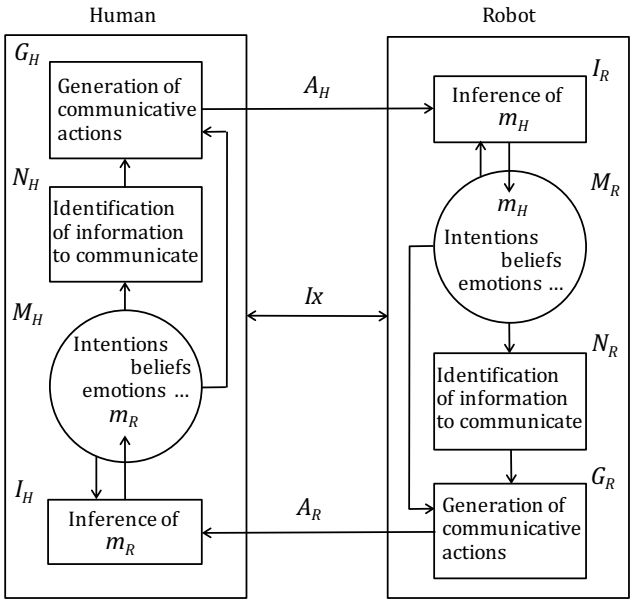

Figure 2: Model of how a human and robot infer an understanding of each other's state-of-mind, through communicative actions A_H and A_R. m_R denotes the human's model of the robot's state-of-mind M_R, and m_H denotes the robot's model of the human's mind M_H. Actions A_R are generated based on the mismatch $|m_R - M_R|$ between M_R and the robot's estimation of m_R (which is part of m_H), and aim at reducing this mismatch. See text for further explanation.

However, this does not mean that m_R necessarily has to be explicitly estimated. The important thing is to estimate the mismatch such that Q1b can be answered.

b) How should it be determined if the mismatch $|m_R - M_R|$ is large enough to generate communicative actions?

c) Which information should be communicated to reduce $|m_R - M_R|$? Sometimes it may be sufficient to simply communicate the fact that there is a significant mismatch. In other cases, the part of M_R causing the mismatch should be communicated. For example, if the mismatch concerns the robot's intended next action, information may be *what* the robot intends to do (Examples 1, 2, 3, 4), *how* it will be done (Example 3), and *why* the robot intends to do it (Example 2). Communicating *why* may be important also for non-action parts of the SoM. As an example, if a social robot behaves in a "stressed manner", the interacting human might benefit from knowing that the reason is that they are running late for a planned bus trip.

d) At which level of details should communication take place? The communication should normally be as brief as possible, while still providing all necessary information. For example, it is sometimes sufficient to communicate an intended goal position by referring to it by name (e.g. "I am going to the kitchen"), but sometimes an exact *x*-*y* position is necessary. Also informing about the *why* and *how* for an action, may involve a trade-off between brevity and sufficiency.

Q2 **(I_R) How should the robot represent and infer the human's mind?**

a) What entities (if any) of the human mind M_H should be represented in the robot's model m_H? These entities are needed to estimate the mismatch $|m_R - M_R|$, and Q2a is therefore tightly connected with question Q1a.

b) How should these entities be represented?

c) How should m_H be inferred from communicative actions A_H performed by the human, from the robot's mind M_R, and from regular interaction Ix?

Q3 **(G_R) How should communicative actions be generated to communicate the required information?** The purpose of communicative actions is to contribute to a change in the human's model m_R such that it becomes more similar to the robot's mind M_R. A large variety of modalities and techniques are possible. The robot may inform the human *explicitly* by dedicated actions (Examples 1, 2, 5, 6, and 7), for example using spoken utterances, joint attention, eye contact, gestures, facial expressions, body language [17, 95], proximity, light projections [44, 89], animated lights [82], or augmented reality [90, 91]. The robot may also inform *implicitly* by adapting its behaviour to convey the necessary information (Examples 3, 4, and 8 above), for example using emotional expressions, paralanguage (e.g. rhythm, intonation, tempo, or stress), or motion variation (e.g. speed or choice of path). It is noteworthy that also a null action may provide information on the SoM, and hence have a communicative function. For example, a robot not turning towards a person who talks to it signals that it is busy with something else.

Outside the scope of the model, two particularly important questions to address are:

Q4 **To whom should the robot direct the communicative actions?** Normally the robot interacts with a single person, but in more complex scenarios, such as a robot navigating among several bystanders, deciding with whom to communicate becomes an important and non-trivial question.

Q5 **Which mechanism should enable the model?** In many cases, the robot should not constantly generate communicative actions in order to reduce the difference between M_R and the robot's estimation of m_R. Rather, a mechanism should identify situations in which it is important for the human to understand the

Brought to you by | Umea University Library
Authenticated
Download Date | 10/26/18 4:01 PM

robot. This often is a non-trivial question to answer, and contains sub questions such as:
a) When should the information be communicated?
b) Should the robot initiate communication or should the robot respond to requests by the human? In the former case, timing may be essential, not least in collaborative settings where mutual understanding of planned actions are essential for both robot and human.

5.1 Example scenarios for the questions

The proposed model may together with the questions above serve as support and inspiration when designing functionality that makes a robot understandable for interacting humans. The remainder of this section gives an example of such a process, using the case described in [44] as baseline and expanding it with alternative solutions and approaches by considering alternative answers to questions Q1-Q5. The overall goal of [44] is defined as having a robotic fork-lift interacting smoothly and safely with humans moving in the same area. In terms of understanding, this means that the human should understand how the robot plans to move (other approaches such as having the robot giving way for humans are not considered here).

Q1 a,b) Given the overall goal, the mismatch $|m_R - M_R|$ concerns the discrepancy between the human's estimation of the planned path of the robot, and the actually planned path of the robot. Q1a and Q1b may be answered in several, fundamentally different ways, and we will in parallel consider two choices: The first choice is denoted TM_0: $|m_R - M_R|$ is assumed to take a constant value, larger than a given threshold for generation of communicative actions. This corresponds to an assumption that the human always has a significantly incorrect estimation of the planned path of the robot. This can be seen as a zeroth-order ToM, and is the choice made in [44]. The second choice is denoted TM_1: $|m_R - M_R|$ is assigned a value larger than the threshold if and only if the planned path of the robot, and the estimated intended path of the human will lead to a collision. The can be seen as the robot inferring that the human has an incorrect estimation of the planned path of the robot, which corresponds to a second-order ToM of the human. Note that this does not require the robot inferring the human's estimation of the planned path of the robot. The reasoning is rather based on an assumption of rationality: since a collision is eminent, the human's estimation is probably incorrect.
Sub question Q1c concerns which information to communicate. For TM_0, the most obvious choice may be the planned path of the robot, which was also the choice made in [44]. Additional information could be the reason *why* the robot plans to go along the planned path. A simpler choice could be to simply communicate the fact that a collision is imminent.
Q1d concerns the level of detail. In [44], a large part of the planned path was chosen to be communicated. For TM_1, an alternative would be to communicate only the predicted location for collision.

Q2 a) Which parts of the human mind M_H the robot should try to infer and represent in m_H depend essentially on the choices made in Q1a,b. For TM_0, the robot infers no part of M_H. For TM_1, the robot needs to infer the human's intended path.
b) The human's intended path may for example be represented as a list of floor coordinates.
c) The human's intended path may be inferred for example by extrapolating the perceived motion pattern of the human.

Q3 A large number of modalities are possible, each one with advantages and disadvantages. In [44], the choice of information to communicate (Q1c) was the planned path of the robot, and the communicative actions were projected light patterns on the floor. If only the fact that a collision is imminent should be communicated, some kind of warning signal like a flashing light or honking would be possible modalities.

Q4 The simplest choice is to direct the communicative actions to no specific person but rather to anyone "listening", in a broadcasting manner. Addressing specific persons may have the advantage of being less invasive, but would require advanced mechanisms for detection of humans, and would also put constraints on the design of communicative actions.

Q5 The simplest choice is to let communicative actions be generated all the time. Other choices could be to only do it when humans are in motion close to the robot.

The example illustrates how a design process of interaction for understanding may benefit from the proposed five questions and associated sub questions. The systematic analysis reveals alternative formulations of what is actually needed for understanding, and of alternative means to accomplish understanding.

Brought to you by | Umea University Library
Authenticated
Download Date | 10/26/18 4:01 PM

6 Summary and conclusions

We used the term *state-of-mind* (SoM) to refer to all relevant information in a robot's cognitive system, for example actions, intentions, desires, knowledge, beliefs, emotions, perceptions, capabilities and limitations of the robot, task uncertainty, and task progress. We further defined *understanding* of a robot as having sufficient knowledge of the robot's SoM to successfully interact with it. The term *communicative actions* was introduced to refer to robot actions aiming at increasing the interacting human's knowledge of the robot's SoM, i.e. to increase the human's understanding of the robot.

A model of interaction for understanding was proposed. The model describes how communicative actions are generated and performed by the robot to reduce the difference between the robot's SoM, and the robot's inferred model of the human's model of the robot's SoM. The human is modeled in a corresponding fashion. The model applies to cases in which both human and robot utilize a first or higher-order ToM to understand each other, and also to simpler cases in which the robot performs static communicative actions in order to support the human's understanding of the robot. Hence, the model may be used to characterize the large body of existing research that, implicitly or explicitly, deal with understandable robots. The model may also serve as inspiration for continued work on understandable robots that truly exploit the possibilities with a ToM working in both human and robot. A specific valuable insight provided by the model is the conceptual separation of information to be communicated, from the means to communicate, i.e. the communicative action.

Implementation of mechanisms for generation, communication, and interpretation of communicative actions can clearly be daunting tasks, but can be guided by addressing the proposed five questions: *What information (if any) should be communicated to the human?*, *How should the robot represent and infer the human's mind?*, *How should communicative actions be generated to communicate the required information?*, *To whom should the robot direct the communicative actions?*, and *Which mechanism should enable the model?*. Some of the reviewed earlier work provide application specific answers to many or all of these questions, but continued work of fundamental nature is encouraged and seen as necessary to reach general solutions for design of understandable robots.

Acknowledgement: The authors wish to thank several anonymous reviewers for their valuable comments and suggestions for improvements, such as highlighting the need to consider differences in sensing and acting capabilities, and the complex mixtures of general interaction and communicative actions.

References

[1] D. Doran, S. Schulz, T. R. Besold, What does explainable AI really mean? A new conceptualization of perspectives, 2017, arXiv:1710.00794 [cs.AI]

[2] A. Chandrasekaran, D. Yadav, P. Chattopadhyay, V. Prabhu, D. Parikh, It takes two to tango: Towards theory of AI's mind, 2017, arXiv:1704.00717

[3] T. Fong, I. Nourbakhsh, K. Dautenhahn, A survey of socially interactive robots: Concepts, design and applications, Technical Report CMU-RI-TR-02-29, Robotics Institute, Carnegie Mellon University, 2002

[4] S. Bensch, A. Jevtić, T. Hellström, On interaction quality in human-robot interaction, In: *International Conference on Agents and Artificial Intelligence (ICAART)*, 2017, 182–189

[5] T. Nomura, K. Kawakami, Relationships between robot's self-disclosures and human's anxiety toward robots, In: *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, IEEE Computer Society, 2011, 66–69

[6] G. Baud-Bovy, P. Morasso, F. Nori, G. Sandini, A. Sciutti, Human machine interaction and communication in cooperative actions, In: S. I. Publishing (Ed.), *Bioinspired Approaches for Human-Centric Technologies*, 2014, 241–268

[7] M. J. Gielniak, A. L. Thomaz, Generating anticipation in robot motion, In: *2011 RO-MAN*, 2011, 449–454

[8] M. Nilsson, S. Thill, T. Ziemke, Action and intention recognition in human interaction with autonomous vehicles, In: *Experiencing Autonomous Vehicles: Crossing the Boundaries between a Drive and a Ride, Workshop in conjunction with CHI2015 (2015)*, 2015

[9] V. M. Lundgren, A. Habibovic, J. Andersson, T. Lagström, M. Nilsson, A. Sirkka, et al., Will there be new communication needs when introducing automated vehicles to the urban context?, In: *Advances in Human Aspects of Transportation*, 2016, 484, 485–497

[10] L. Wang, G. A. Jamieson, J. G. Hollands, Trust and reliance on an automated combat identification system, *Human Factors*, 2009, 51(3), 281–291

[11] M. M. de Graaf, B. F. Malle, A. Dragan, T. Ziemke, Explainable robotic systems, In: *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '18, New York, NY, USA, ACM, 2018, 387–388

[12] L. Takayama, D. Dooley, W. Ju, Expressing thought: Improving robot readability with animation principles, In: *2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2011, 69–76

[13] C. Lichtenthäler, T. Lorenzy, A. Kirsch, Influence of legibility on perceived safety in a virtual human-robot path crossing task, In: *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, 2012, 676–681

[14] C. Lichtenthäler, *Legibility of Robot Behavior Investigating Legibility of Robot Navigation in Human-Robot Path Crossing Sce-*

Brought to you by | Umea University Library
Authenticated
Download Date | 10/26/18 4:01 PM

narios, PhD thesis, Technische Universität München, 2014
[15] K. Dautenhahn, S. Woods, C. Kaouri, M. Walters, K. L. Koay, I. Werry, What is a robot companion - friend, assistant or butler?, In: *Proc. IEEE IRS/RSJ Int. Conference on Intelligent Robots and Systems*, Edmonton, Alberta, Canada, 2005, 1488–1493
[16] A. D. Dragan, K. C. Lee, S. S. Srinivasa, Legibility and predictability of robot motion, In: *Proceedings of the 8th ACM/IEEE International Conference on Human-robot Interaction*, HRI '13, Piscataway, NJ, USA, 2013, 301–308 IEEE Press
[17] J. Novikova, *Designing Emotionally Expressive Behaviour: Intelligibility and Predictability in Human-Robot Interaction*, PhD thesis, University of Bath, 2016
[18] H. Karvonen, I. Aaltonen, Intent communication of highly autonomous robots, In: *The 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2017)*, 2017
[19] N. Mirnig, M. Tscheligi, Comprehension, coherence and consistency: Essentials of robot feedback, In: J. Markowitz (Ed.), *Robots that talk and listen – technology and social impact*, De Gruyter, 2015
[20] J. Knifka, On the significance of understanding in human-robot interaction, In: M. Nørskov (Ed.), *Social Robots: Boundaries, Potential, Challenges*, Ashgate, 2016, 3–17
[21] C. Breazeal, Towards sociable robots, *Robotics and autonomous systems*, 2002, 42
[22] K. Dautenhahn, The art of designing socially intelligent agents: Science, fiction, and the human in the loop, *Applied Artificial Intelligence*, 1998, 12(7-8), 573–617
[23] R. H. Wortham, A. Theodorou, J. J. Bryson, Robot transparency, trust and utility, *Connection Science*, 2017, 29(3), 242–248
[24] J. B. Lyons, Being transparent about transparency: A model for human-robot interaction, In: *Proceedings of AAAI Spring Symposium on Trust in Autonomous Systems*, 2013, 48–53
[25] M. P. Anderson, What is communication, *The Journal of Communication*, 1959, 5(9)
[26] C. E. Shannon, A mathematical theory of communication, *Bell System Technical Journal*, 1948, 3(27), 379–423
[27] D. Chandler, The transmission model of communication, http://visual-memory.co.uk/daniel/Documents/short/trans.html, 1994, (Accessed: May 20 2018)
[28] W. Schramm, *The Beginnings of Communication Study in America*, Thousand Oaks, CA: Sage, 1997
[29] C. L. Baker, J. B. Tenenbaum, Modeling human plan recognition using bayesian theory of mind, In: *Plan, activity, and intent recognition: Theory and practice*, 2014, 177–204
[30] S. Baron-Cohen, *Mindblindness: An essay on autism and theory of mind*, MIT Press, Cambridge, 1995
[31] D. G. Premack, G. Woodruff, Does the chimpanzee have a theory of mind?, *Behavioral and Brain Sciences*, 1978, 1(4), 515–526
[32] M. Michlmayr Simulation theory versus theory theory: Theories concerning the ability to read minds, Master's thesis, University of Innsbruck, 2002
[33] P. M. Churchland, Folk psychology and the explanation of human behavior, In: J. D. Greenwood (Ed.), *The future of folk psychology*, Cambridge University Press, Cambridge, 1991, 51–69
[34] R. Verbrugge, L. Mol, Learning to apply theory of mind, *Journal of Logic, Language and Information*, 2008, 4(17), 489–511
[35] B. Hare, J. Call, M. Tomasello, Do chimpanzees know what conspecifics know and do not know?, *Animal Behaviour*, 2001, 61(1), 139–151
[36] T. Bugnyar, S. A. Reber, C. Buckner, Ravens attribute visual access to unseen competitors, *Nature Communications*, 2015, 7
[37] L. M. Hiatt, J. G. Trafton, A cognitive model of theory of mind, In: *International Conference on Cognitive Modeling*, 2010
[38] L. Barlassina, R. M. Gordon, Folk psychology as mental simulation, In: E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Summer Edition)*, 2017
[39] M. A. Arbib, A. Billard, M. Iacoboni, E. Oztop, Synthetic brain imaging: Grasping, mirror neurons and imitation, *Neural Networks*, 2000, 13, 975–997
[40] A. I. Goldman, Theory of mind, In: E. Margolis, R. Samuels, S. P. Stich (Ed.), *The Oxford Handbook of Philosophy of Cognitive Science*, 2012
[41] P. Carruthers, Simulation and self-knowledge: a defence of theory-theory, In: P. Carruthers, P. R. Smith (Ed.), *Theories of theories of mind*, Cambridge University Press, Cambridge, 1996, 22–38
[42] G. Gergely, Z. Nfidasdy, G. Csibra, S. Biro, Taking the intentional stance at 12 months of age, *Cognition*, 1995, 56, 165–193
[43] E. Bonchek-Dokow, *Cognitive Modeling of Human Intention Recognition*, PhD thesis, Bar Ilan University, 2012
[44] R. Chadalavada, H. Andreasson, R. Krug, A. J. Lilienthal, That's on my mind! robot to human intention communication through on-board projection on shared floor space, In: *Proceedings of European Conference on Mobile Robots*, 2015
[45] S. Augustsson, J. Olsson, L. G. Christiernin, G. Bolmsjö, How to transfer information between collaborating human operators and industrial robots in an assembly, In: *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, NordiCHI '14, New York, USA, ACM, 2014, 286–294
[46] M. Matthews, G. V. Chowdhary, Intent communication between autonomous vehicles and pedestrians, In: *Robotics: Science and Systems*, 2015
[47] K. Kobayashi, S. Yamada, Making a mobile robot to express its mind by motion overlap, In: V. A. Kulyukin (Ed.), *Advances in Human-Robot Interaction*, InTech, 2009
[48] D. Dennett, *The Intentional Stance*, MIT Press, Cambridge, 1987
[49] F. Hegel, S. Krach, T. Kircher, B. Wrede, G. Sagerer, Theory of mind (ToM) on robots: A functional neuroimaging study, In: *HRI'08*, Netherlands, 2008, 335–342
[50] X. Zhao, C. Cusimano, B. F. Malle, Do people spontaneously take a robot's visual perspective?, In: *HRI '16: The Eleventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, 2016
[51] S. lai Lee, I. Y. man Lau, S. Kiesler, C.-Y. Chiu, Human mental models of humanoid robots, In: *Proceedings of the 2005 IEEE International Conference on Robotics and Automation ICRA 2005*, IEEE, 2005, 2767– 2772
[52] T. Nakata, T. Sato, T. Mori, Expression of emotion and intention by robot body movement, In: *Proc. of the Intl. Conf. on Autonomous Systems*, 1998
[53] A. Zhou, D. Hadfield-Menell, A. Nagabandi, A. D. Dragan, Expressive robot motion timing, In: *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, HRI 2017, Vienna, Austria, March 6-9 2017*, 2017, 22–31
[54] T. Ono, M. Imai, R. Nakatsu, Reading a robot's mind: A model of utterance understanding based on the theory of mind mechanism, *Advanced Robotics*, 2000, 14(4), 311–326

Brought to you by | Umea University Library
Authenticated
Download Date | 10/26/18 4:01 PM

[55] S. H. Huang, D. Held, P. Abbeel, A. D. Dragan, Enabling Robots to Communicate their Objectives, *ArXiv e-prints*, 2017
[56] H. Kautz, *A Formal Theory of Plan Recognition*, PhD thesis, University of Rochester, 1987
[57] E. Charniak, R. P. Goldman, A Bayesian model of plan recognition, *Artificial Intelligence*, 1993, 64(1), 53–79
[58] M. Vilain, Getting serious about parsing plans: A grammatical analysis of plan recognition, In: *Proceedings of National Conference on Artificial Intelligence*, 1990
[59] H. H. Bui, S. Venkatesh, G. West, Policy recognition in the abstract hidden Markov model, *Journal of Artificial Intelligence Research*, 2002, 17, 451–499
[60] E. A. Billing, T. Hellström, A formalism for learning from demonstration, *Paladyn, Journal of Behavioral Robotics*, 2010, 1(1), 1–13
[61] E. A. Billing, T. Hellström, L. E. Janlert, Behavior recognition for learning from demonstration, In: *Proceedings of IEEE International Conference on Robotics and Automation*, Alaska, Anchorage, 2010, 866–872
[62] A. Billard, S. Calinon, R. Dillmann, S. Schaal, *Robot Programming by Demonstration*, Springer, 2008
[63] C. L. Nehaniv, K. Dautenhahn, *Of hummingbirds and helicopters: An algebraic framework for interdisciplinary studies of imitation and its applications*, World Scientific Press, 2000, 24, 136–161
[64] B. Jansen, T. Belpaeme, A computational model of intention reading in imitation, *Robotics and autonomous systems*, 2005, 54, 394–402
[65] S. Trott, M. Eppe, J. Feldman, Recognizing intention from natural language: Clarification dialog and construction grammar, In: *Proceedings of 2016 Workshop on Communicating Intentions in Human-Robot Interactionnd Systems in NYC, NY*, 2016
[66] A. Sutherland, S. Bensch, T. Hellström, Inferring robot actions from verbal commands using shallow semantic parsing, In: H. Arabnia (Ed.), *Proceedings of the 17th International Conference on Artificial Intelligence ICAI'15*, 2015, 28–34
[67] A. Rasouli, L. Kotseruba, J. K. Tsotsos, Agreeing to cross: How drivers and pedestrians communicate, In: *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, 2017
[68] B. Scassellati, Theory of mind for a humanoid robot, *Autonomous Robots*, 2002, 12, 13
[69] A. M. Leslie, Tomm, toby, and agency: Core architecture and domain specificity, In: l. A. Hirschfeld, S. A. Gelman (Ed.), *Mapping the Mind: Domain Specificity in Cognition and Culture*, Cambridge University Press, Cambridge, 1994, 119–148
[70] B. Benninghoff, P. Kulms, L. Hoffmann, N. Krämer, Theory of mind in human-robot-communication: Appreciated or not?, *Kognitive System*, 2013, 1
[71] M. Berlin, J. Gray, A. L. Thomaz, C. Breazeal, Perspective taking: An organizing principle for learning in human-robot interaction, In: *Nat. Conf. on Artificial Intelligence, vol. 21*. AAAI Press, MIT Press, 2006
[72] G. Milliez, M. Warnier, A. Clodic, R. Alami, A framework for endowing an interactive robot with reasoning capabilities about perspective-taking and belief management, In: *Int. Symp. on Robot and Human Interactive Communication*, IEEE, 2014, 1103–1109
[73] K.-J. Kim, H. Lipson, Towards a simple robotic theory of mind, In: *Proceedings of the 9th Workshop on Performance Metrics for Intelligent Systems (PerMIS09)*, New York, USA, ACM, 2009, 131–138
[74] S. Devin, R. Alami, An implemented theory of mind to improve human-robot shared plans execution, In: *The Eleventh ACM/IEEE International Conference on Human Robot Interaction (HRI16)*, Piscataway, NJ, USA, IEEE, 2016, 319–326
[75] L. M. Hiatt, A. M. Harrison, J. G. Trafton, Accommodating human variability in human-robot teams through theory of mind, In: *IJCAI International Joint Conference on Artificial Intelligence*, 2011, 2066–2071
[76] C. Bereiter, *Education and mind in the Knowledge Age*, L. Erlbaum Associates, 2002
[77] D. Vernon, S. Thill, T. Ziemke, The role of intention in cognitive robotics, In: A. Esposito, L. C. Jain (Ed.), *Toward Robotic Socially Believable Behaving Systems - Volume I*, pages 15–27. Springer International Publishing, 2016
[78] S. Thrun, J. Schulte, C. Rosenberg, Robots with humanoid features in public places: A case study, *IEEE Intelligent Systems archive*, 2000, 15(4), 7–11
[79] F. Stulp, J. Grizou, B. Busch, M. Lopes, Facilitating intention prediction for humans by optimizing robot motions, In: *International Conference on Intelligent Robots and Systems (IROS)*, 2015
[80] H. Romat, M.-A. Williams, X. Wang, B. Johnston, H. Bard, Natural human-robot interaction using social cues, In: *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, HRI '16, Piscataway, NJ, USA, IEEE Press, 2016, 503–504
[81] J. Hough, D. Schlangen, It's not what you do, it's how you do it: Grounding uncertainty for a simple robot, In: *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI '17)*, 2017
[82] K. Baraka, M. M. Veloso, Mobile service robot state revealing through expressive lights: Formalism, design, and evaluation, *International Journal of Social Robotics*, 2017
[83] B. Kühnlenz, S. Sosnowski, M. Buß, D. Wollherr, K. Kühnlenz, M. Buss, Increasing helpfulness towards a robot by emotional adaption to the user, *Int J Soc Robot*, 2013, 5(4), 457–476
[84] A. Moon, B. Panton, M. V. der Loos, E. Croft, Using hesitation gestures for safe and ethical human-robot interaction, In: *IEEE ICRA'10 Workshop on Interactive Communication for Autonomous Intelligent Robots*, 2010, 11–13
[85] R. A. Knepper, On the communicative aspect of human-robot joint action, In: *IEEE International Symposium on Robot and Human Interactive Communication Workshop: Toward a Framework for Joint Action, What about Common Ground?*, New York, NY, USA, 2016
[86] R. A. Knepper, C. I. Mavrogiannis, J. Proft, C. Liang, Implicit communication in a joint action, In: *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, pages 283–292, New York, NY, USA, ACM, 2017
[87] A. Sciutti, G. Sandini, Interacting with robots to investigate the bases of social interaction, *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2017, 25, 2295–2304
[88] C. Breazeal, A. Edsinger, P. Fitzpatrick, B. Scassellati, Active vision systems for sociable robots, *IEEE Trans. Syst. Man Cybern.*, 2001, 31, 443–453
[89] A. Watanabe, T. Ikeda, Y. Morales, Communicating robotic navigational intentions, In: *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015
[90] R. T. Azuma, A survey of augmented reality, *Presence*, 1997, 6(4), 355–385

Brought to you by | Umea University Library
Authenticated
Download Date | 10/26/18 4:01 PM

[91] J. Carff, M. Johnson, E. M. El-Sheikh, J. E. Pratt, Human-robot team navigation in visually complex environments, In: *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2009)*, 2009, 3043–3050

[92] C. Breazeal, P. Fitzpatrick, That certain look: Social amplification of animate vision, In: *AAAI 2000 Fall Symposium*, 2000, 18–22

[93] F. Broz, A. Di Nuovo, T. Belpaeme, A. Cangelosi, Talking about task progress: Towards integrating task planning and dialog for assistive robotic services, *Paladyn, Journal of Behavioral Robotics*, 2015, 6(1), 111–118

[94] R. Kelley, A. Tavakkoli, C. King, M. Nicolescu, M. Nicolescu, Understanding activities and intentions for human-robot interaction, In: D. Chugo (Ed.), *Advances in Human-Robot Interaction*, In-Tech, 2010, 288–305

[95] H. Knight, R. Simmons, Layering laban effort features on robot task motions, In: *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, New York, NY, USA, ACM, 2015, 135–136

[96] J. G. Trafton, N. L. Cassimatis, M. D. Bugajska, D. P. Brock, F. E. Mintz, A. C. Schultz, Enabling effective human-robot interaction using perspective-taking in robots, *Systems, Man and Cybernetics*, 2005, 35(4), 460–470

Brought to you by | Umea University Library
Authenticated
Download Date | 10/26/18 4:01 PM