International Journal of Social Robotics (2023) 15:1227–1241
https://doi.org/10.1007/s12369-023-01022-z

A Mini-survey on Psychological Pillars of Empathy for Social Robots: Self-Awareness, Theory of Mind, and Perspective Taking

Elahe Bagheri[1]

Accepted: 15 May 2023 / Published online: 1 July 2023

Abstract
Recent studies in the field of Human–Robot Interaction (HRI) confirm the positive effects of robots' empathic behaviors in HRI. Most HRI studies investigating empathy, apply an empirical approach to implement empathy, i.e., the empathic model is derived directly from observations of empathic actions. This resulted in the emergence of numerous different empathic models that are only valid for a particular scenario that is highly tuned and, therefore, a slight modification in the scenario makes the corresponding empathy model infeasible. In fact, most of the proposed models suffer from a lack of generalizability. Since empathy is a complex concept that includes different dimensions, a coherent model of empathy that can be used in different scenarios or even be scenario independent, needs to consider several core concepts of empathy.Thus, the goal of this paper is to analyze and link different concepts of empathy and bring related existing models together, which can help researchers in the HRI community to have a better picture of an empathy model that might lead to the development of more general models of empathy for social robots.

Keywords Empathy · Self-awareness · Theory of mind · Perspective taking · Cognitive architecture · Human–robot interaction · Social robot

1 Introduction

Nowadays, social robots are investigated to care of elderlies in elder homes [1, 2], help them to live longer independently [3], assist in education [4, 5], serve as tutors at schools [6], reinforce social behaviors in children with autism [7], help in rehabilitation tasks [8], promote shopping [9], and serve as guides in museums [10, 11]. As robots' interactions with humans in real dynamic environments is increasing, they need to be able to interact naturally with humans. One way to enable robots to establish a natural interaction is by developing human norms in them[12]. Among different proposed robot behavior models, those that adjust their behavior based on humans' social norms are more preferred [13].

The idea of developing robots that explicitly show human social behaviors emerged in the early 1990 s. Bartneck and Forlizzi [14] defined a *social robot* as: "an autonomous or semi-autonomous robot that interacts and communicates with humans by following the behavioral norms expected by the people with who the robot is intended to interact". Thus, besides accomplishing assigned tasks, social robots should also be able to interact and communicate with humans in a socially appropriate manner.

One common way to improve interaction between humans and robots is to first study Human–Human Interaction (HHI) characteristics, and later apply them to social robots. Although in some cases this is not successful, e.g., "uncanny valley", which shows HRI is not the same as HHI and often needs particular consideration, in many cases the HHI results extend properly to the HRI cases. For instance, humans with different personality types have different preferences, e.g., introvert people mostly prefer talking with lower volume, lower speed, and prefer praising comments, while extrovert people mostly prefer talking louder, faster, and challenging comments. Thus, Tapus and Mataric [15] and Esteban et al. [16] applied the effects of personality in HHI into a social robot's behavior model and evaluated it in an HRI scenario to verify whether considering personality for social robots improves their interaction with humans.

In another case, Ivaldi et al. [17] and Cao et al. [18] considered HHI engagement approaches to make social robots more

✉ Elahe Bagheri
elahe.bagheri@ivai.onl

[1] IVAI GmbH, Syke, Germany

engageable. In fact, when disengagements happen, humans try to adapt their behavior to regain others' engagement. Ivaldi et al. [17] and Cao et al. [18] applied these human behaviors (verbal and non-verbal) to social robots to enable them to regain users' engagement.

Another example of applying HHI characteristics to HRI is determining robot's social distance. Giddings [19] argued social acceptance can be seen as a process in which people evaluate, generate, reevaluate, and refine their social distance from others. Similarly, Kim and Mutlu [20] argued that humans might engage in a similar process with robots and Human–Robot social distance might serve as a multidimensional construct that shapes people's acceptance of robots. Another example of mapping HHI characteristics to HRI studies is emotion expression. Sacks et al. [21] showed that expressing emotions is expected at certain conditions in an interaction and Fischer et al. [22] argued emotional expression plays a considerable social role in the regulation of interpersonal relationships, which led Fischer et al. [22] to enhance a robot's emotion expression behavior by following human social norms. In another study, to enable robots to apply empathic behaviors towards humans, findings by De Vignemont and Singer [23] about how and when humans' empathic behaviors can be more accepted by other humans, are used to determine the level of robot's empathic behavior [24].

Previous studies revealed that applying HHI norms and characteristics mostly improves the interaction between humans and robots. In addition, it is shown that, among all HHI norms and characteristics that are applied to HRI, robots that show empathy are considered as more acceptable, likely, trustworthy, supportive [25], friendly [26], engageable [27], and have a higher chance that humans make long-term interactions with them [28].

Empathy is one of the major elements in humans' social interactions [29] by which humans assess another person's situational context and then respond to it by expressing empathic behaviors [30]. Accordingly, once a robot understands the emotional state of another person, it can change its behavior to adjust it to the other's affective state and express empathic behaviors.

To develop a coherent empathic model we need to understand the concept of empathy, however, empathy is an interdisciplinary concept that is studied in different fields such as psychology [29], neuroscience [31], and philosophy [32]. In this paper, the main focus is on explaining the psychological components of empathy. The most related psychological concepts to empathy are self-awareness, Theory of Mind (ToM), and perspective taking, such that Asada [33] argued empathy may only occur in animals with self-awareness, and both affective and cognitive empathy (Sect. 2) require a distinction between one's own and others' mental states and a representative form of one's own embodied emotions.

Regard to the relation of empathy and ToM, Baron-Cohen et al. [34] and Meltzoff [35] stated that children with autism have deficits in ToM and empathy, and Meltzoff [36] said that an infant's ability to imitate others lies at the origins of ToM, perspective taking, and empathy. In respect to the importance of perspective taking, Goldstein and Winner [37] showed that activities that need stepping into others' shoes, i.e., perspective taking, lead to growth in both empathy and ToM.

Due to the relation between these concepts, they are even used interchangeably, for instance, Hynes et al. [38] defined empathy as an emotional perspective taking, and ToM as a cognitive perspective taking, and Baron-Cohen and Wheelwright [39] and Blair [40] used ToM as synonymous with cognitive empathy. Baron-Cohen et al. [41], Gillberg [42], Kaland et al [43] and Roeyers et al. [44] used ToM interchangeably with empathy and Kalbe et al. [45] used ToM instead of empathy. Charlop–Christy and Daneshvar [46] broke down ToM into an operationally defined behavior of perspective taking, and Maurage et al. [47] used cognitive empathy as synonymous with perspective taking. However, Davis [48] has considered each concept as an individual component of empathy and highlighted differences between them. Following Davis, this paper also analyzes each concept as an individual component and introduces them from a psychological point of view. In addition, to illustrate how these concepts can be developed and later combined to propose a general empathy model, the state-of-the-art models that tried to develop these concepts in the field of HRI are reviewed.

This paper is structured as follows: Sect. 2 discusses the concept and definition of empathy from a psychological point of view. Section 3 focuses on self-awareness and starts with explaining the concept of self-awareness in psychology and then explaining proposed models of self-awareness, reviewing related work, and finally proposing methodologies to include self-awareness in an empathy model. Section 4 defines and outlines the concept of ToM, reviews its related state-of-the-art models, and proposes possible approaches for integrating ToM into an empathy model. Section 5, similar to the two previous sections, describes perspective taking, explains its types, and reviews its related state-of-the-art models. Since the reviewed models of self-awareness, ToM, and perspective-taking, are focused only on one topic and not integrating these concepts, in Sect. 6 characteristics of a comprehensive model of empathy are discussed and cognitive architectures, which aim to model humans' minds, are reviewed to investigate their potential usage in developing a general model of empathy for social robots. Finally, Sect. 7 concludes this paper.

2 Definition of Empathy

2.1 Definition and Construct of Empathy

Empathy is a complex component with many different definitions in psychology, for instance, Cuff et al. [50] identified 43 distinct definitions for empathy. Originally, empathy has been considered as either a cognitive or an affective phenomenon. Empathy as a cognitive phenomenon is the process where the observer, i.e., empathizer, can understand what the other person, i.e., target, is experiencing by taking her perspective and detecting her internal state but without necessarily experiencing any emotional change. Thereby, the empathizer can provide some reactions more congruent with the target's feeling than her own feeling [51]. Hodges and Myers [52] argued cognitive empathy is more like a skill, in which humans learn to recognize and understand the target's emotional state and respond to it appropriately. On the other hand, empathy as an affective phenomenon, which is also known as "emotional empathy", is an unintentional and uncontrollable process, where the empathizer not only can understand what the target is experiencing but also can feel her emotions by sharing or experiencing her emotional state [49]. While emotional empathy might be unpleasant for the empathizer because of the personal distress and discomfort that happens to her by observing the target's negative feelings and conditions [53], cognitive empathy leads to less personal distress for the empathizer and more concern for the target [54].

The relation between cognitive and affective empathy is not clear in the literature, for instance, Feshbach [55] considered cognitive empathy as a prerequisite for affective empathy, Eisenberg and Strayer [56] believes both emotional and affective dimensions of empathy are directly related, and Hoffman [29] believes both types of empathy work together to produce an empathic response. Some researchers also suggest that being able to recognize and understand others' emotions, i.e., cognitive empathy, is a necessary but not sufficient component of affective empathy.

However, Davis treated empathy as a multidimensional phenomenon that includes both cognitive and affective components [49] (Fig. 1). He defined empathy as a set of constructs that connect the responses of the empathizer to the experiences of the target. These constructs include both the "processes" taking place within the empathizer and the affective and non-affective "outcomes" that result from these processes. The main constructs in his prototype are:

- **Antecedents**, which refer to the attributes of the empathizer, target, or situation;
- **Processes**, which refer to the process by which empathic outcomes are formed;
- **Intrapersonal outcomes**, which refer to the cognitive, affective, and motivational empathic outcomes that are formed in the empathizer but are not necessarily shown to the target;
- **Interpersonal outcomes**, which refer to the behavioral empathic outcomes that are shown to the target.

2.2 Output of Empathy

Some researchers like Duan and Hill [57] believe that empathic output emotions should be the same as emotions that are experienced by the target. However, equality of the expressed emotions by the empathizer to those of the target does not necessarily lead to a positive effect for the target, e.g., if the target is sad and the empathizer expresses only sadness, it will not necessarily decrease the target's sadness, as Costa et al. [58] showed, if the target is sad due to injustices, people may express anger, which shows more empathy towards the target than being upset. In fact, it is possible that developing different emotions (from those of the target) makes the target eventually less sad. Davis labeled this simple matching of emotions for responding or reacting to the target's feeling as parallel empathy and introduced reactive empathy as a reaction that goes beyond this and tries to comfort the target by expressing different emotional states than what the target is experiencing [59].

Through parallel empathy, the empathizer mimics the target's emotions by synchronizing facial and vocal expressions, postures, and movements with those of the target,

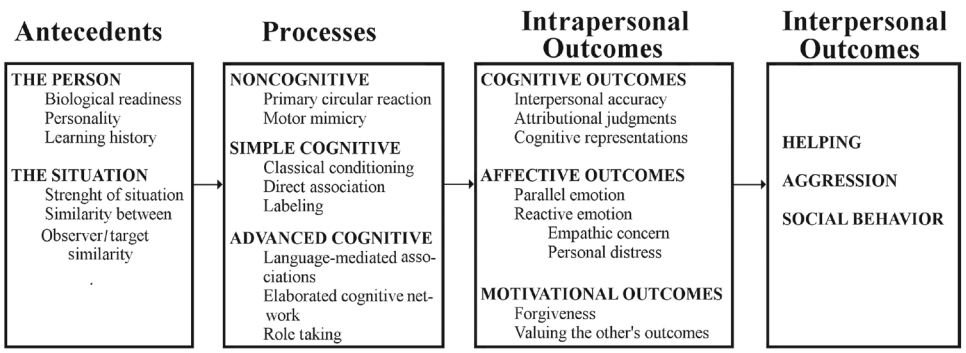

Fig. 1 Proposed prototype of empathy by Davis [49], which considers both cognitive and affective outcomes as a part of empathy (this figure is duplicated from [49])

which can be seen as emotional contagion [60]. Emotional contagion refers to a process in which the emotions or behaviors of a person or a group are influenced by another person's or group's emotional states and behavioral attitudes [61]. Dimberg and Thunberg [62] argued people who express emotional empathy are more strongly susceptible to empathic contagion. This is possible by mirror neurons, which are fired both when one "acts" or "observes" another one is acting the same [63, 64], i.e., whether one sees another's emotional state or consciously adopts his/her psychological view, similar neural circuits are activated in the self. Different researchers argued that the mirror neuron system is involved in empathy [65–68].

The reactive outcomes of empathy, on the other hand, aim to alter and enhance the target's affective state. Different studies defined reactive empathy as an emotional response that is unlike what the target is experiencing [69–71]. Todd and Galinsky [72] believe reactive empathy involves a complementary emotional reaction reflecting concern for a target's well-being. While parallel outcomes of empathy are more self-oriented [71], reactive outcomes are focused on the target and require more advanced cognitive processing, e.g., perspective taking, thus, the reactive outcome can be considered as a higher level of empathic behavior [26, 70, 73].

2.3 Modulation Factors on Empathy

De Vignemont and Singer [23] introduced four main categories of factors that modulate human's empathic behaviors as the following:

- **Intrinsic features of the shared emotion**: the intensity, saliency, and valence (positive versus negative) of the emotion that is expressed by the target modulates the empathic behavior expressed by the empathizer.
- **Characteristics of the empathizer**: the type of empathic behavior that empathizer expresses is modulated by empathizer's gender [39], personality[1] [74], age [75], and past experiences [76].
- **Relationship between the empathizer and the target**: the kind of relationship between the empathizer and the target, e.g., competitive or cooperative relationship [77].
- **Situational context**: the context in which the empathizer observes the target. For instance, if the empathizer is confronted with several targets who display different emotions or if the reasons of the expressed emotion by the target are not clear.

Further, some studies investigated the effect of other factors like humor on empathy. For instance, Hampes [78] found a positive correlation between empathy and affiliative[2] and self-enhancing[3] humor. On the other hand, he found a negative correlation between empathy and self-defeating[4] and aggressive[5] humor.

In addition, Wang et al. [79] showed that expressing empathy and humor improves the interaction between a virtual agent and students in the context of e-learning.

This section outlined the definition of empathy, its main components, outputs, and modulation factors. In the following sections, the three pillar concepts of empathy that have been introduced in the introduction, are investigated individually.

3 Self-Awareness and Empathy

Self-awareness is the ability to reflect on one's own cognition [80], which can be in different levels, e.g., being aware of our body, which enables us to recognize ourself in the environment, being aware of our mental states, which enables us to know our feelings, desires, and beliefs [81], or being able to monitor and follow our thought process (self monitoring), which is a metacognitive skill, and enables us to regulate our strategies (self-organizing) [82].

Self-awareness prohibits the overlap between self and other representations and prevents confusion between self and other's feelings, which can induce emotional distress or anxiety [83]. By self-awareness, we can consciously know and understand our own character, feelings, motives, and desires. Disability in self-awareness can lead to personal distress, i.e., a self-focused and averse response to another's emotional state, and hampers the ability to toggle between self and other's perspectives.

Different studies showed that having self-awareness can improve the efficiency of a robot in open, complex, and dynamic environments [84], however, there is no unique definition of the way self-awareness could be integrated into a robot's behavior [85]. Some researchers believe that even if a robot has no complete self-awareness, it can have some characteristics of self-awareness such as the ability to recognize itself in a mirror, being aware of its own health status, or having emotional states. For example, Michel et al. [86] developed an infant-like humanoid robot called Nico that can recognize its moves in its visual field as well as in a mirror. Nico expects to see a motion in its visual field, whenever certain motor movements commence, after a certain time. Thereby, it can distinguish itself from others based on the

[1] The effect of targets' personality on their preferred empathic behaviors is also investigated in different studies, e.g., [15, 24, 51].

[2] Telling jokes and saying funny things.

[3] Making humor in stressful or adversity situations of life.

[4] Allowing others to make humor about oneself, and laughing along with others when being ridiculed or disparaged.

[5] Using humor to attack or tease other people.

idea of linking motion to time. Another sample of body-discovery is [87], where Bongard et al. developed a model that enables a robot to continuously create a concept of its own physical structure. The proposed self model is used to generate forward movements for a four-legged machine that uses actuation sensation relationships to infer its own structure indirectly. If the robot's physical structure changes unexpectedly, e.g., a leg part is removed, it can rebuild its internal self model to produce new behaviors to cope with these changes, e.g., generating alternative gaits. In a more recent work, Saegusa et al.[88] used vision and proprioception sensory inputs to enable the applied platforms, i.e., iCube and James, to define their own body parts. To achieve this, the correlation between the motions in the vision field and proprioception is calculated to verify whether the moving object in the visual field is related to its own motor function or not. Once an object is determined as correlated to motor activity, it is considered as its own body, and data related to the body posture and other visumotor parameters are stored in a memory. This way robot recognized its own body parts.

While Michel et al. [86] and Bongard et al. [87] focused on physical body self-awareness, Steinfeld et al. [89] and Anshar and Williams [90] used self-awareness to enable the robot to have an overview on its physical condition. For instance, Steinfeld et al. [89] argued that self-awareness is important to evaluate whether involving a human partner for assisting a robot is useful or not, i.e., if the robot is not aware of its capabilities and is not able to recognize its troubles, it requires a human for monitoring and intervention. Similarly, Anshar and Williams [90] believed that robots need to be aware of their internal state of well-being, since if a robot is damaged it may put people around it at risk of injury, and by self-awareness can prevents it by informing and warning its human collaborator. They interpreted the robot's damage as its pain and as the concept of pain in humans is strongly related to the concept of self-awareness, thus, Anshar and Williams [90] proposed a robot design framework, i.e., adaptive self-awareness framework (ASAF), to evolve appropriate self-awareness and pain concepts for robots to enable them to be aware of their damages. To this end, ASAF has five different components, i.e., consciousness, synthetic pain description, robot mind, action execution, and database. Robot consciousness is the cognitive aspect of the robot that specifically signifies the focus of the robot's attention. Synthetic pain description simulates a synthetic pain by setting some joint restriction regions that the robot should avoid. The robot mind allows it to adapt to the world by predicting its own future states through reasoning about the perceived/detected facts. The action execution module executes one of the three decisions that the robot can make, i.e., sending an alert (to inform the human about its damage), modifying the joint stiffness values (to repair the damage), or shifting the awareness of the robot about its body parts, which prevents further impact on the robot's hardware and possible harm to the human partners in the case of robot damage.

However, Novianto and Williams [84] argued, if a robot can only recognize its own motion or itself in a mirror cannot be considered as a self-aware robot since this kind of recognition capabilities can be obtained via a specific program and does not necessarily need a genuine awareness capacity. Instead, they considered a robot to be self-aware, if it can focus its attention on the representation of its internal states, e.g., emotions, intentions, and beliefs. To achieve this, they designed ASMO (Attentive Self Modifying) framework, which provides concepts like perception, attention, and self modification, and offers a mechanism for directing and creating behaviors. To determine what is happening in the system and the world, ASMO has twofold facing, i.e., outward and inward facing, while the former senses physical stimuli outside the robot's body, the latter senses inside physical stimuli. Processing the inward and outward sensations, perceptions are created. Based on the created perceptions and also the provided attention and self modification mechanisms, ASMO enables the robot to deliberate and re-plan its behaviors. The model is evaluated in a scenario in which a humanoid robot is playing an instrument and a human comes and takes the instrument from the robot. In response, an unhappy feeling is evoked in the robot, and it starts crying and asking the human to give the instrument back. Meanwhile, the robot's attention may be directed to this stimulus (unhappy feeling), in this case, it realizes that crying and requesting the instrument does not lead to getting the instrument back. Thus, self modification mechanism provides two other reactions for the robot. These reactions are either stopping crying and asking the human to give the instrument back or informing the human that it has finished playing and wants to do something else. In fact, ASMO simulates cognition as a set of autonomous independent processes, where each process has an attention value, which is either directly assigned or learned from experience. Attention values vary dynamically and affect robot's actions [91]. Later, Novianto [92] updated ASMO by adding an attention mechanism, which mediates the competition between processes, an emotion mechanism, which biases the amount of attention is demanded by different processes, and a learning mechanism, which adapts robot's attention to improve its performance.

Another work towards developing mental self-awareness model is Kawamura et al. [93] where sense of self is represented in the self-agent, which contains self-reflection, self-awareness, and sense-of-self. Self-agent is the location of planning systems, executive control, self-monitoring, and task selection. It is continually updated and enhanced to allow the robot to reason and act based on its status and the context of its tasks. Self-agent consists of a set of simple agents interacting with memory systems. The memory

structure is divided into three classes: Short-Term Memory (STM), Long-Term Memory (LTM), and Working Memory System (WMS). STM holds sensory information about the current environment, while the LTM learns and teaches behaviors, experiences, and semantic knowledge. The WMS holds task-specific STM and LTM information and streamlines the information flow to the cognitive processes during the task. By the implemented self-agent, the robot can deliberate its emotions based on memory experience. The emotion that emerges through an experience is learned and stored in the memory systems. Later, when a new event occurs, evoked emotions activate the episodic memory, which in return activates cognitive control to suppress current behavior and execute required behavior [94].

Although none of the previous developed models utilized self-awareness for expressing empathy, to have a general model of empathy, all these individual abilities are important and necessary. For instance, it is important that robots be able to understand their internal state, what is necessary to do, and the consequence of actions they take on their future status. In addition, robots need to be able to change their attention from the task they are doing to their human partners' states. Also, it is important to have a model, which enables the robot to be aware of its hardware status, since if the robot needs to move to a target to show empathy and it has some disabilities due to a hardware damage (or lack of electrical power) it should reconsider its empathic behavior. To achieve all this, a model of self-awareness is required that records all attitudes of the robot, e.g., past experiences, internal states and hardware condition.

4 Theory of Mind and Empathy

The previous section explained the role of self-awareness in empathy and showed how being able to distinguish your mind from others is important to develop empathy. Next, ToM, which refers to the assumption that others also have a mind similar to one's own [95], is required. ToM has three orders, first order states that everyone has a belief of him/her self, e.g., A thinks..., the second order is about having a model of another's mind, i.e., A thinks that B thinks..., and the third order refers to when A has a model of how B is thinking about A or C, e.g., I know what you are thinking I am thinking [96].

Having ToM allows us to understand and attribute feelings, desires, intentions, and thoughts to others and informs us that others act according to their feelings and intentions [97], which can be used to explain and predict their behaviors [95]. For instance, when an empathizer is observing a target, ToM enables her to model the target's mental state and predict her reactions. Dvash and Shamay–Tsoory [98] argued ToM is a part of a person's empathic ability and is typically involved in generating cognitive empathic responses, such that a deficit in ToM can lead to a decreased cognitive empathic response, and Holopainen et al. [99] showed that training ToM, i.e., performing exercises like emotion recognition, pretense, false belief, and humor improves the empathy ability of children with autism, who suffer a deficit in ToM. These studies confirm the correlation between ToM and empathy.

The importance of ToM in empathy comes from the fact that through ToM one can predict and understand other's internal states and feelings, which are crucial for empathy. However, Goldstein et al. [100] showed that strength in ToM can exist independently of strength in empathy, as actors are skilled in ToM but they do not express empathy more than average in comparison to others, and Winter et al. [101] found that aggressive offenders who showed reduced empathic responses to emotional videos of others' suffering, had an intact performance of ToM. To be able to do (at least complex forms of) empathy, ToM is necessary but having ToM does not lead to empathy necessarily. On the other hand, Salazar Kämpf et al. [102] compared abilities of ToM and empathy in two groups of healthy individuals and people with obsessive-compulsive disorder (OCD), who exhibited higher levels of empathy in comparison to healthy individuals. Obtained results show that although people with OCD express a higher level of empathy, concerning ToM, no differences are detected between the two groups, which shows that stronger abilities of empathy do not necessarily need stronger abilities of ToM.

ToM is mainly studied by two theories, namely "theory-theory" and "simulation-theory". Theory-theory asserts that humans hold a basic or naive theory of psychology to infer the mental states of others such as their beliefs, desires, or emotions. This information is then used to understand the intentions behind others' actions or predict their future behavior [103]. This theory supports the affective component of empathy stronger than simulation-theory [45].

On the other hand, simulation-theory holds that humans anticipate and make sense of others' behaviors by activating mental processes that, if carried into action, would produce similar behaviors. For instance, children use their own emotions to predict what others will do [104]. In fact, simulation-theory states that certain parts of the brain have dual use, such that they are not only used to generate our own behaviors and mental states but also to predict and infer others' behaviors and mental states [105]. These findings fit neatly with the mirror neuron's findings, which state that behaviors can be simulated by activation of the same neural resources for acting and perceiving [106]. Simulation-theory uses more biological evidence [107] and better supports the cognitive component of empathy [108].

Although in this paper the focus is on the relation between ToM and empathy, developed models of ToM in HRI, have focused mainly on applying perspective taking and

belief management abilities in robots, and there is no work using ToM to develop empathy. However, summarizing the reviewed papers, two types of works are mostly performed for modeling ToM in HRI. The first type tries to show how ToM is working, e.g., [109] and [110], where in [109] the effect of robotic appearance is investigated on evoking ToM in humans and in [110] interaction of agents endowed with ToM is investigated. In the second type of works, i.e., [111] and [112], the advantages of endowing a robot with a model of ToM, in different situations, are investigated. Following, these works are described in more detail.

Riek et al. [109] investigated the effect of robotic factors on evoking ToM in humans. To this end, a 30 s film clip featuring five protagonists of varying degrees of human-likeness are shown to participants to see how people make empathy with them. Results showed people are more empathic towards human-like robots and less empathic towards mechanical-like robots, which is compatible with the simulation-theory that states people mentally simulate the situation of others to understand their mental and emotional state, such that the more robots are human-like, the better humans can project their situations into their own mental states. Additionally, the results showed that the more the robot is human-like, the stronger the expressed empathy is perceived, which supports findings by Krach et al. [113] who argued people view anthropomorphize robots as more like themselves. Unfortunately, the effect of other factors like the robot's gender, age, size, language, background culture, etc. is not investigated. Also, it would be interesting to investigate the effect of endowing the robot with ToM on the behavior of people towards it, i.e., does seeing a robot with ToM change people's behavior towards it?

Similarly, Devin and Alami [111] tried to use ToM to enable a robot to understand the mental state and intention of its interactant and adjust its behavior towards her. To achieve this, Devin and Alami [111] proposed a framework which consists of six different modules, including (a) a situation assessment module, which evaluates the world's current state from all agents' point of view based on the spatial perspective taking (Sect. 5), (b) a high-level task planner, which allows the robot to synthesize shared plans containing the actions of all agents involved in a given task, (c) a supervisor module, which manages the execution of the shared plans, (d) a geometric action and motion planner, for computing trajectories as well as objects' placements and grasps to perform actions, (e) a dialogue manager to verbalize information to the human and to recognize basic vocal commands, and finally (f) a ToM module, which takes the models computed by the situation assessment module and also the status of goals, plans, and actions from the supervisor module to estimate and maintain the mental state of each agent involved in the cooperation. Thereby, the robot knows if the human's mental state is not up to date.

The scenario in which the model is tested is a "clean the table" scenario in which a robot and a human have to clean a table together by first, removing all items from it, second, sweeping it, and third, replacing all items on it. The objects on the table are either reachable only by one of the agents or by both of them. If the human removes all the objects that only she can remove and then leaves the room or starts talking on the phone, the robot continues the task and removes the rest of the objects, sweeps the table, and puts the items (that are reachable for it) back on the table. When the human partner comes back, she sees some items that she did not move are still on the table and she may think that the robot had stopped working after she left and the table is not swept yet. However, as the robot is able to estimate her mental state, it can update her about the current state of the world, and prevent her from sweeping the table.

In another work, Peters [112] used ToM to understand others' interest in interacting with a robot. To this end, he proposed a model consisting of different modules including synthetic vision, visual attention, direction of attention detector, mutual attention detector, and theory of mind. Through these modules, he investigated users' interaction characteristics like greeting gestures, gaze, head, and body direction to obtain their interest level in an interaction. The proposed model has been tested in a virtual world and showed that the applied ToM module is able to determine an agent's interest level in interaction, and coordinates the other agent's behaviors accordingly. However, humans' behaviors may change in different situations or even in the same situations but at different times. To cope with these variations, Hiatt et al. [114] used ToM to identify what different beliefs, desires, or intentions can lead to different behaviors in similar situations. To figure it out, Hiatt et al. [114] designed a patrolling task in which the robot has two main approaches for selecting a path: first, using a probabilistic simulation analysis, and second, using a hypothetical generation model. The former analyzes the simulation to see what different paths can be observed by executing the probabilistic model and assigns each path a probability. With this information, the robot is able to find the most likely execution path that matches the human's action. The second approach, i.e., the hypothetical generation model, is used when simulation analysis does not explain the change in human behavior, i.e., the robot asks the human why she is doing what she is doing and memorizes her answer. Next time the human does something unexpected, it checks whether the newly learned knowledge led her to behave differently.

Previous works showed that considering ToM improves the interaction between humans and robots by enabling the robot to adjust and coordinate its behavior with the human partner. However, in all these experimental settings, the applied models of ToM aim to model others' mental states regarding the defined task and as the goal of the robot and

its interactant is the same, e.g., both want to avoid a collision or sweep a table, building a model of the others' minds for the specified task comes down to modeling their information about current task status. However, to apply empathy, a robot should have a more general model of the other's mind, which not only covers what the target is (apparently) focusing on but also covers their affective state and their reactions. For instance, while a human and a robot are sweeping a table, if the human leaves the room and comes back with a different emotional state, which can be observed in her facial expression, speech, or body language, the robot's model of her mind should ensure the robot that this change is not related to the current interaction between them but, most likely, an external stimuli.

In addition, based on such a model, the robot should be able to predict what would be her reaction, if the robot tries to start empathizing with her, considering different empathic behaviors. In fact, only by such a model, a robot is able to perform empathy in the right moment and in the right fashion.[6]

Mainly, to endow a robot with a simple form of ToM, two steps are necessary to be taken. First, the robot needs to understand its user's mental state. This can be achieved by reasoning on the robot's contextual information and sensory input data, e.g., visual and auditory inputs, so that the robot can predict the user's mental state and feelings. In the second step, the robot needs to analyze the user's mental state to predict her goals and intentions and her potential reaction to achieve them. Although this is challenging, assuming having an accurate model of the user's mind, affective parameters on the user, and knowledge of the environment, the robot can predict the user's next actions, either by looking into previous similar situations or by reasoning about effects of the current stimuli on the user.

5 Perspective taking and Empathy

The last two sections emphasized the importance of (a) having the ability to distinguish your mind from others (self-awareness) and (b) having a model of others' minds (ToM). This section describes the importance of being able to be in the other's shoes, which is known as perspective taking.

Perspective taking is the process by which one sees a situation from another's point of view, which has been shown to strengthen both parallel and reactive empathy [115–117]. Perspective taking has been defined along two dimensions: perceptual and conceptual [118]. The perceptual dimension describes the ability to understand how other people experience things through their senses, e.g., visually or auditory [118]. The literature of the perceptual dimension, has mostly focused on the visual perspective taking, i.e., the ability to understand the way another person sees things in physical space[7] [120]. Visual perspective taking has been applied in different domains, e.g., to improve the accuracy of activity recognition and recognizing a human's actions [121], to resolve ambiguities in an operator's command [122], to learn a task from ambiguous demonstrations [105], and to approach a target while hiding from sight [123].

The conceptual dimension, on the other hand, focuses on the ability to comprehend and take the viewpoint of another person's psychological experience, i.e., thoughts, feelings, and attitudes [118]. Conceptual perspective taking is used to simulate the decision making process of others to predict their next action in competitive [124] and cooperative [125] scenarios.

Following, two types of recent developed models of perspective taking are reviewed, first, models that use perspective taking to enable a robot to adapt to its user's behavior, i.e., [119, 126, 127] and [128], and second, a model that uses perspective taking to manipulate human's actions, i.e., [129].

Lemaignan et al. [119] used perspective-taking to enable a robot to be aware of geometric reasoning and situation assessment of its environment, i.e., the robot knows different capabilities from the perspectives of another agent, e.g., what the other agent can see, what the other agent is focused on, and which object is pointed to by the other agent. In a shared task with a human partner, this knowledge helps the robot to correctly interpret what the human says, and to plan tasks that the human partner is able to do. In this manner, the robot can successfully share space and tasks with a human partner.

Fischer and Demiris [126] equipped iCub robot with a depth camera to be able to perform two different types of visuospatial perspective taking. To this end, the robot needs to first, learn the environment, second, recognize objects within the environment, third, estimate the gaze and head pose of the surrounding humans, and finally, determine whether an object is visible for a human partner. The model, also enables the robot to estimate what the world looks like to the human. To do so, the environment is mapped in the reference frame of the human and is then mentally rotated. In fact, through mental perspective transformation, the world is reconstructed from another viewpoint. The model is verified through a scenario in which a human asks the robot to grasp an object, although the robot can see two objects, through perspective taking, it understands that only one of them is in the human's sight and therefore instead of asking which object to grasp, grasps the intended one.

[6] In general, the more the robot knows about the target, the more appropriate, accurate, and personalized will be its empathic behaviors.

[7] Another type of perceptual perspective taking is spatial perspective taking, which refers to the qualitative spatial location of objects (or agents) with respect to a frame [119].

Similarly, Pandey and Alami [127] presented an affordance graph, which contains both agent-object, and agent-agent perspectives, and shows what an agent is able to do with an object, and also what it can do for another agent. To achieve this, the proposed model contains different graphs, e.g., taskability, manipulability, and affordance graph. The taskability graph encodes what all agents in the environment might be able to do for all other agents, with which levels of mutual efforts and at which places. The manipulability graph encodes what an agent might be able to do with an object, and with which effort level. In fact, while the taskability graph encodes agent-agent affordances, the manipulability graph represents agent-object affordances. By combining a set of taskability graphs and a set of manipulability graphs for a set of affordances, the concept of an affordance graph is developed, which reveals the action-possibilities of manipulating the objects among the agents and across different places, and also shows information about the required level of effort and the potential spaces. Hence, the affordance graph enables an agent to determine the action capabilities of other agents. To examine the proposed model, a scenario is defined in which a robot along with two human partners tries to pick objects on a table. Meanwhile, humans not only move and change their positions but they also change the position of the objects on the table. Applying the proposed model, the robot is able to update its model of the world and determine achievable objects for different users, and also understand the possible actions of different users dynamically.

In another work, a simulation-theory based model is proposed to enable a robot to understand the environment from the perspective of social partners to infer the intention of their instruction and, once the robot finds the human's intentions, it focuses only on the important subset of the problem space, which helps the robot to learn a task. In this order, Berlin et al. [128] emphasized the importance of perspective taking in the concept of teaching new tasks to robots by demonstration, i.e., a robot needs to understand a human teacher's perspective to learn from her demonstration. To enable the robot to understand the world from its own perspective and the teacher's perspective, two individual components are proposed for each one, i.e., a perception system which, represents the world from robot's perspective and a belief system, which represents the world from the teacher's perspective. The perception system extracts perceptual features from raw sensory information and generates the robot's beliefs. To generate the human teacher's beliefs, the belief system clusters the perceptual information into discrete object representations by considering spatial relationships between the various observations and in conjunction with other metrics of similarity. During a learning episode, the robot records the states of its own perception system and teacher's belief system to infer the goal from observed differences in these two worlds during this episode. To evaluate the proposed model, a general assembly task is designed in which the human teacher tries to teach the robot to put a peg in the object's hole. However, one of the objects is behind a barrier such that the robot can see it but the teacher cannot, thus, the teacher does not put a peg in this object's hole. Yet, using the proposed model, the robot can take the teacher's perspective and understand that this object is out of her sight, otherwise the same rule was applied to it and a peg was placed in its corresponding hole.

In a more advanced form of perspective taking, Breazeal [129] proposed a model to manipulate a human user's mental state through the robot's physical actions. To this end, the robot obtains a model that shows how a chosen action changes the world and how the changed world changes the user's mental state. Using this model, the robot is able to take the user's perspective and perform actions that manipulate the users' mental state in order to achieve its goals. To examine the proposed model a competitive game was designed, where a human and a robot have to take an object from point A and put it in point B. The robot wins, if the two players place different objects, and the human wins if the objects are the same. While points A and B are hidden for the other player, they can see each other on the way from A to B.

Three scenarios are defined to examine whether the robot can manipulate the user's mental state. In the first condition, the robot aims to hide the main object it wants to play and meanwhile leads the opponent to believe that it is carrying another object, thereby, it carries the decoy object openly while it caries the main object behind itself. In the second condition, the robot only wants to hide the main object from the human, therefore, it caries the object behind itself, and finally, in the third condition, the robot transports the object while the opponent has a 50% chance to see the object.

The obtained results showed that the proposed perspective taking model enabled the robot to manipulate the human's mental state, i.e., in the first condition, the human selected the decoy object, in the second condition, a random object, and in the third condition, the object that could be observed.

Previous works showed that endowing robots with perspective taking, enables them to better understand their human interactant's intention and reason of their behavior, which not only smooths their interaction but also decreases ambiguity in the interaction. Yet, most of the works are focused on visual tasks, where creating a model of the world enables the robot to visit it from different view points and obtain how others see the world, which can enable the robot to perform some forms of empathy. For instance, in a scenario where the human is upset/angry because of loosing a personal item (which is not lost but is out of her sight, and the robot can see it) the robot can use the proposed models by Berlin et al. [128] and Fischer and Demiris [126] to take the human's perspective and apply reactive empathy by showing her the object. However, for having a general model of

empathy, not only visual perspective is important, but mental perspective also is important, because there are situations in which users may not necessarily want to show their feelings or intentions.

To predict others' real point of view, two approaches exist, one, using our model of others' mind (second order of ToM) to imaging them in that situation and putting ourselves in their shoes (taking their perspective), second, using our own model of world (first order of ToM), in the case there exist no model of others' minds, and imagining our self in their situation [130]. However, due to individual differences, the results of the former approach might be different than the latter. Yet, using perspective taking enables us to infer others' feelings, intentions, and reason of their actions in the current situation. And indeed, the more accurate and comprehensive our model of others' minds (ToM) is, the better we obtain their perspective of the world in the current frame of the world. Similarly, once the robot has a good model of others' minds, taking their perspective becomes quite straightforward.

6 Discussion

Empathy is a complex phenomenon that is the result of the interaction between different cognitive abilities (Fig. 2). The previous sections shed light on the role of different cognitive abilities involved in expressing empathy, i.e., self-awareness, ToM, and perspective taking. To express any form of empathy (even mimicking the target's affective state, which is the simplest form of empathy), a self-awareness module is required by which the robot is able to distinguish the target's feelings from its own (Fig. 2). This module should enable the robot to find the data related to itself among all its sensory input, e.g., its hardware status and abilities, its knowledge about others, etc.

In addition, the empathy model should be able to find data related to the target, e.g., her facial expression, speech, body language, etc. and data related to the surrounding environment, e.g., data related to her dog sitting in the other corner of the room, or the movie she is watching on the TV. In addition, the empathy model needs the target's model of mind to analyze sensory data from her perspective to find her attentions and emotions and reason about them to understand the meaning of the current sensory data, e.g., is she crying because she is sad or happy?, is she angry because of the movie or because her dog has broken her vase? Only by such a model it is possible to understand what led the target to the current state, what are her current intentions and goals and what can be done to change her affective state to a better one, if necessary.

Further, the empathy model should be able to analyze the effect of showing any forms of empathy on the target's future emotional state. This helps the model to provide the most appropriate empathic behavior towards the target, which can be parallel or reactive and convey similar or different emotions. This ability also is achievable by having the target's model of the mind, which enables the model to predict the target's reactions to different empathic behaviors and evaluates the effectiveness of each proposed behavior before expressing it towards the target so that after analyzing its consequences, an adjusted version of this empathic behavior be expressed towards the target (Fig. 2).

And finally, the model should be able to express the proposed empathic behavior via facial expressions, speech, body gestures, etc. The way different modalities are combined and used to express the robot's empathic behavior, depends on different parameters including robot's abilities, target's age, culture, personality, strength and type of the target's emotion. This procedure should continue until a final state is achieved, e.g., target feels better, robot goes out of resources, or target asks the robot to stop empathizing.

Therefore, a general model of empathy that can be used by social robots (or any other artificial agent) needs to have four fundamental modules that fulfill the following requirements:

- **Sensory input**, which collects visual, auditory (verbal and non-verbal), tactile, gustatory, somatosensory, and any other form of input data.
- **Mental States**,[8] which represent different types of contents, e.g., cognitive contents such as beliefs, intentions, goals, memories (episodic, semantic, procedural), as well as affective contents such as emotional states.
- **Mapping of sensory input to Mental States**,[9] finding the current state of the mind, which can be done via perception and attentional mechanisms, e.g.,[131].
- **Mapping of Mental states to Actions**, finding appropriate action in current mental state, which can be done through consciously accessible or automatic mechanisms.

To achieve any general enough model of empathy, it is necessary to develop all these four abilities not only for self-modeling but also for hetero-modeling, i.e., modeling of others' minds. In fact, for higher levels of empathy, where one needs to understand and estimate other's mental state and then consciously select an action that will change the other person's mental state towards a better state, a representation of the other person's mind is required which enables the one to select the appropriate actions. Thus a "models of minds" is required, which is often referred to as "cognitive architectures". An example of a cognitive architecture that aims to model human cognition at the process level is ACT-R (Adaptive Character of Thought-Rational)[10][132]. ACT-R consists

[8] Can be considered as ToM.

[9] Can be considered as Perspective-taking.

[10] http://act-r.psy.cmu.edu/.

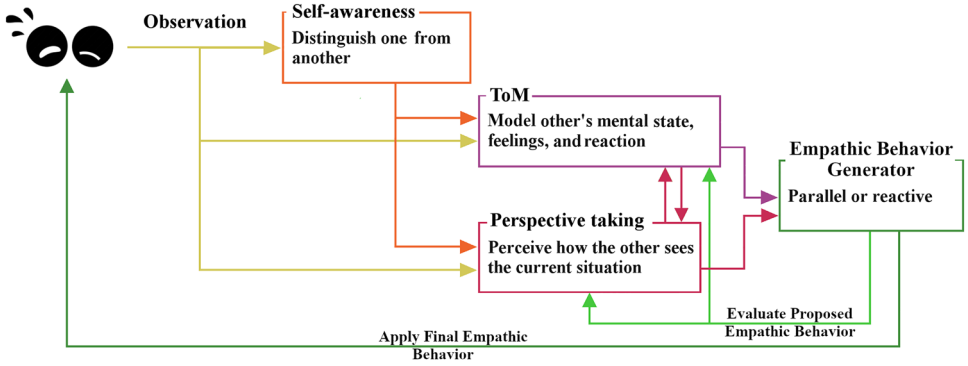

Fig. 2 The relation between different psychological concepts of empathy and the approaches that can be used and combined to generate empathic behaviors. To perform empathy, a self-awareness module, which enables the robot to distinguish between its own feelings and other's feelings, a model of ToM, and perspective taking by which the robot can understand how the target is feeling and predict her emotional state are required. In addition, the robot can evaluate its empathic behavior through perspective taking and ToM modules to estimate the target's reaction to the proposed empathic behavior and adjust it, if needed, before expressing it to the target

of different modules including visual, aural, vocal, manual, imaginal, goal, declarative, and a central system as procedural. Each module is associated with a specific brain region and has a role, e.g., the aural module is able to search its auditory environment and recognize sounds and utterances, the visual module observes elements in the model's world, the imaginal module holds external information, the manual module saves connections to the outside world, the goal module holds control states, the declarative module stores facts and critical information, and the procedural system executes steps of a procedure. Each module has a small capacity known as buffer that stores a small amount of data to represent the current attention of the corresponding module. The contents of the buffers at a given moment in time, represents the state of the ACT-R at that moment.

However, Birlo and Tapus [85] argued a simple interaction between the robot's external states in terms of in and outputs via the external world is not sufficient for the robot to be able to act self-aware, instead, a robot has to create its own interpretation of what it perceives and connect this information to its current internal state as well as to its previous states. To achieve this, they focused on the representation of internal states. They used ACT-R and added a meta-cognition module to it, which represents the self and is an independent unit that looks over all buffers and memory contents, and decides on which of these buffers it will pay attention to. The self module has access to every module and buffer of ACT-R/E (Adaptive Character of Thought-Rational/Embodied) to be able to retrieve information about memories and possible actions, and also the current robot's working memory. Based on all this information, the self module determines the content of its self buffer. The content of the self buffer represents the focus of the system's attention on a meta-level. By having all the other buffer contents as well as ACT-R/E's current focus of attention "in mind", the self is able to interfere with what is happening inside ACT-R/E's procedural module. The procedural module determines the system's behavior and sets the current focus of attention. As the self module has the capability to interfere in the processes of the procedural module, it can deliberate and re-plan ACT-R/E's behavior.

Later, Trafton et al. [133] adapted ACT-R by adding two new modules to enable spatial reasoning in a three-dimensional world and made modifications to perceptual and motor modules to allow the tight linkage between perception and action to function in the embodied world by placing an additional constraint on cognition, i.e., cognition occurs within a physical body that must perceive the world, navigate and maneuver in space, and manipulate objects.

Although existing cognitive architectures are not used for expressing empathy, they have modules required for developing an empathy model, e.g., self-awareness, reasoning, perception detection, attention detection, etc, and can be used as inspiration for developing general models of empathy.

7 Conclusion

This paper provides a brief and summarized overview of the psychological background of empathy, which can be used by HRI researchers to develop a more comprehensive model of empathy. The types and levels of empathy are explained, and its related psychological concepts are discussed. The corresponding concepts, i.e., self-awareness, ToM, perspective taking, and also cognitive architectures as a mechanism for developing a model of the mind, are discussed individually. In addition, the most accepted definition for each concept, the relation between them, and their use cases are also outlined. Further, the most recent developed models for each

concept, in the field of HRI, are reviewed and explained in the corresponding Sections.

To the best of my knowledge, no model uses self-awareness, ToM, or perspective taking, to apply empathy, i.e., corresponding HRI models are detached and each model is focused on the corresponding concept and is examined in a specific scenario to verify the proposed model. To fill this gap, the current paper, emphasized on the importance of incorporating these concepts to build a comprehensive model of empathy and discussed potentials of cognitive architecture to achieve this. In fact, having a cognitive architecture that includes self-awareness, ToM, and perspective taking, is able to solve different challenges in applying empathy, e.g., finding the right meaning of the expressed emotion, finding the most effective empathic behavior, finding the appropriate time for applying empathy, and finally, ability to evaluate the applied empathic behavior.

Acknowledgements I would like to thank my colleague Oliver Roesler for our discussions and his valuable comments. I would also like to thank the reviewers for their helpful comments.

Data Availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current paper.

Declarations

Conflict of interest The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Ethical Approval No participant was involved in the process of the article, and the article is not analyzing any user data.

References

1. Kachouie R, Sedighadeli S, Khosla R, Chu M-T (2014) Socially assistive robots in elderly care: a mixed-method systematic literature review. Int J Human Comput Interact 30(5):369–393
2. Bemelmans R, Gelderblom GJ, Jonker P, De Witte L (2012) Socially assistive robots in elderly care: a systematic review into effects and effectiveness. J Am Med Dir Assoc 13(2):114–120
3. Van Kemenade Margo AM, Konijn Elly A, Hoorn Johan F (2015) Robots humanize care. In: Proceedings of the international joint conference on biomedical engineering systems and technologies, Vol 5. pp 648–653. SCITEPRESS-Science and Technology Publications, Lda
4. Belpaeme T, Kennedy J, Baxter P, Vogt P, Krahmer EEJ, Kopp S, Bergmann K, Leseman P, C Küntay A, Göksun T et al (2015) L2tor-second language tutoring using social robots. In: Proceedings of the ICSR 2015 WONDER Workshop
5. Vogt Paul, de Haas Mirjam, de Jong Chiara, Baxter Peta, Krahmer Emiel (2017) Child-robot interactions for second language tutoring to preschool children. Front Human Neurosci. https://doi.org/10.3389/fnhum.2017.00073
6. Belpaeme Tony, Kennedy James, Ramachandran Aditi, Scassellati Brian, Tanaka Fumihide (2018) Social robots for education: a review. Sci Robot. https://doi.org/10.1126/scirobotics.aat5954
7. Coeckelbergh M, Pop C, Simut R, Peca A, Pintea S, David D, Vanderborght B (2016) A survey of expectations about the role of robots in robot-assisted therapy for children with ASD: ethical acceptability, trust, sociability, appearance, and attachment. Sci Eng Ethics 22(1):47–65
8. Matarić MJ, Eriksson J, Feil-Seifer DJ, Winstein CJ (2007) Socially assistive robotics for post-stroke rehabilitation. J Neuroeng Rehabil 4(1):5
9. De Gauquier L, Cao HL, Gomez EP, De Beir A, van de Sanden S, Willems K, Brengman M, Vanderborght B (2018) Humanoid robot pepper at a belgian chocolate shop. In: Companion of the 2018 ACM/IEEE international conference on human–robot interaction
10. Burgard W, Cremers AB, Fox D, Hähnel D, Lakemeyer G, Schulz D, Steiner W, Thrun S (1999) Experiences with an interactive museum tour-guide robot. Artif Intell 114(1–2):3–55
11. Yamazaki A, Yamazaki K, Ohyama T, Kobayashi Y, Kuno Y (2012) A techno-sociological solution for designing a museum guide robot: regarding choosing an appropriate visitor. In: 2012 7th ACM/IEEE international conference on human-robot interaction (HRI)
12. Scheutz M, Schermerhorn P, Kramer J, Anderson D (2007) First steps toward natural human-like hri. Auton Robot 22(4):411–423
13. Dautenhahn K, Walters M, Woods S, Koay KL, Nehaniv CL, Sisbot A, Alami R, Siméon T (2006) How may i serve you? A robot companion approaching a seated person in a helping context. In: Proceedings of the 1st ACM SIGCHI/SIGART conference on Human–robot interaction, pp 172–179
14. Bartneck Christoph, Forlizzi J (2004) A design-centred framework for social human-robot interaction. In: RO-MAN 2004. 13th IEEE International workshop on robot and human interactive communication (IEEE Catalog No. 04TH8759), pp 591–594. IEEE
15. Tapus Adriana, Mataric MJ (2008) Socially assistive robots: the link between personality, empathy, physiological signals, and task performance. In: AAAI spring symposium: emotion, personality, and social behavior, pp 133–140
16. Esteban PG, Bagheri E, Elprama SA, Jewell Charlotte IC, Cao HL, De Beir A, Jacobs A, Vanderborght B (2021) Should i be introvert or extrovert? A pairwise robot comparison assessing the perception of personality-based social robot behaviors. Int J Soc Robot, pp 1–11
17. Ivaldi S, Lefort S, Peters J, Chetouani M, Provasi J, Zibetti E (2017) Towards engagement models that consider individual factors in hri: on the relation of extroversion and negative attitude towards robots to gaze and speech during a human-robot assembly task. Int J Soc Robot 9(1):63–86
18. Cao HL, Torrico Moron PC, Esteban GP, De Beir A, Bagheri E, Lefeber D, Vanderborght B (2019) HMM, did you hear what i just said? Development of a re-engagement system for socially interactive robots. Robotics 8(4):95
19. Giddings FH (1911) The relation of social theory to public policy. Am J Sociol 16(5):577–592
20. Kim Y, Mutlu B (2014) How social distance shapes human-robot interaction. Int J Hum Comput Stud 72(12):783–795
21. Sacks H, Schegloff EA, Jefferson G (1978) A simplest systematics for the organization of turn taking for conversation. In: Studies in the organization of conversational interaction, Elsevier, pp 7–55
22. Fischer K , Jung M, Jensen LC, aus der Wieschen MV (2019) Emotion expression in hri–when and why. In: 2019 14th ACM/IEEE international conference on human–robot interaction (HRI)
23. De Vignemont F, Singer T (2006) The empathic brain: How, when and why? Trends Cogn Sci 10(10):435–441
24. Bagheri E, Esteban PG, Cao HL, De Beir A, Lefeber D, Vanderborght B (2020) An autonomous cognitive empathy model

responsive to users' facial emotion expressions. ACM Trans Interact Intell Syst 10(3):1–23. https://doi.org/10.1145/3341198
25. Brave S, Nass C, Hutchinson K (2005) Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. Int J Hum Comput Stud 62(2):161–178
26. Rasool Z, Masuyama N, Islam MdN, Loo CK (2015) Empathic interaction using the computational emotion model. In: 2015 IEEE symposium series on computational intelligence, pp 109–116. IEEE
27. Leite I, Martinho C, Paiva A (2013) Social robots for long-term interaction: a survey. Int J Soc Robot 5(2):291–308
28. Leite I, Castellano G, Pereira A, Martinho C, Paiva A (2014) Empathic robots for long-term interaction. Int J Soc Robot 6(3):329–341
29. Hoffman ML (2001) Empathy and moral development: implications for caring and justice. Cambridge University Press, Cambridge
30. McQuiggan SW, Lester JC (2006) Learning empathy: a data-driven framework for modeling empathetic companion agents. In: Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems, pp 961–968. ACM
31. Minoru A (2015) Development of artificial empathy. Neurosci Res 90:41–50
32. Nakao H, Itakura S (2009) An integrated view of empathy: psychology, philosophy, and neuroscience. Integr Psychol Behav Sci 43(1):42
33. Asada M (2015) Towards artificial empathy. Int J Soc Robot 7(1):19–33
34. Baron-Cohen S (2001) Theory of mind in normal development and autism. Prisme 34(1):74–183
35. Simon BC, Leslie Alan M, Uta Frith (1985) Does the autistic child have a theory of mind? Cognition 21(1):37–46
36. Meltzoff AN (2002) Imitation as a mechanism of social cognition: origins of empathy, theory of mind, and the representation of action. Blackwell handbook of childhood cognitive development, pp 6-25
37. Goldstein TR, Winner E (2012) Enhancing empathy and theory of mind. J Cogn Dev 13(1):19–37. https://doi.org/10.1080/15248372.2011.573514
38. Hynes Catherine A, Baird Abigail A, Grafton Scott T (2006) Differential role of the orbital frontal lobe in emotional versus cognitive perspective-taking. Neuropsychologia 44(3):374–383
39. Baron-Cohen S, Wheelwright S (2004) The empathy quotient: an investigation of adults with asperger syndrome or high functioning autism, and normal sex differences. J Autism Dev Disord 34(2):163–175
40. Blair J, Robert R (2005) Responding to the emotions of others: dissociating forms of empathy through the study of typical and psychiatric populations. Conscious Cognit 14(4):698–718
41. Baron-Cohen S, Wheelwright S, Hill J, Raste Y, Plumb I (2001) The reading the mind in the eyes test revised version: a study with normal adults, and adults with asperger syndrome or high-functioning autism. J Child Psychol Psychiatry Allied Disciplines 42(2):241–251
42. Gillberg CL (1992) The emanuel miller memorial lecture 1991: autism and autistic-like conditions: subclasses among disorders of empathy. J Child Psychol Psychiatry 33(5):813–842
43. Kaland N, Møller-Nielsen A, Callesen K, Mortensen EL, Gottlieb D, Smith L (2002) A newadvanced'test of theory of mind: evidence from children and adolescents with asperger syndrome. J Child Psychol Psychiatry 43(4):517–528
44. Roeyers H, Buysse A, Ponnet K, Pichal B (2001) Advancing advanced mind-reading tests: empathic accuracy in adults with a pervasive developmental disorder. J Child Psychol Psychiatry 42(2):271–278
45. Kalbe E, Grabenhorst F, Matthias Brand J, Kessler RH, Markowitsch HJ (2007) Elevated emotional reactivity in affective but not cognitive components of theory of mind: A psychophysiological study. J Neuropsychol 1(1):27–38
46. Charlop-Christy MH, Daneshvar S (2003) Using video modeling to teach perspective taking to children with autism. J Posit Behav Interv 5(1):12–21
47. Maurage P, Grynberg D, Noël X, Joassin F, Philippot P, Hanak C, Verbanck P, Luminet O, de Timary P, Campanella S (2011) Dissociation between affective and cognitive empathy in alcoholism: a specific deficit for the emotional dimension. Alcoholism Clin Exp Res 35(9):1662–1668
48. Davis MH (1983) Measuring individual differences in empathy: evidence for a multidimensional approach. J Pers Soc Psychol 44(1):113
49. Davis MH (2006) Empathy. In: Stets J, Turner J (eds) Handbook of the sociology of emotions. Springer, Berlin, pp 443–466
50. Cuff BMP, Brown SJ, Taylor L, Howat DJ (2016) Empathy: a review of the concept. Emot Rev 8(2):144–153
51. Bagheri E, Roesler O, Cao HL, Vanderborght B (2021) A reinforcement learning based cognitive empathy framework for social robots. Int J Soc Robot 13:1079–1093. https://doi.org/10.1007/s12369-020-00683-4
52. Hodges SD, Myers MW (2007) Empathy. Encycl Soc Psychol 1:297–298
53. Wise PS, Cramer SH (1988) Correlates of empathy and cognitive style in early adolescence. Psychol Rep 63(1):179–192
54. Davis Mark H et al (1980) A multidimensional approach to individual differences in empathy
55. Feshbach ND (1990) 12 parental empathy and child adjustment/maladjustment. Empathy Development. p 271
56. Eisenberg N, Strayer J (1987) Critical issues in the study of empathy
57. Duan C, Hill CE (1996) The current state of empathy research. J Couns Psychol 43(3):261
58. Costa S, Brunete A, Bae B-C, Mavridis N (2018) Emotional storytelling using virtual and robotic agents. Int J Humanoid Rob 15(03):1850006
59. Davis M (1996) Empathy: a social psychological approach 1994. Brown and Benchmark Publishers, Madison
60. Hatfield E, Cacioppo JT, Rapson RL (1994) Emotional contagion: Cambridge studies in emotion and social interaction. In: Cambridge, UK: Cambridge University Press. errors-in-variables regression model when the variances of the measurement errors vary between the observations. Statistics in Medicine, 21:1089–1101
61. Schoenewolf G (1990) Emotional contagion: behavioral induction in individuals and groups. Mod Psychoanal 15(1):49–61
62. Dimberg U, Thunberg M (2012) Empathy, emotional contagion, and rapid facial reactions to angry and happy facial expressions. PsyCh J 1(2):118–127
63. Keysers C (2009) Mirror neurons. Curr Biol 19(21):R971–R973
64. Rizzolatti G (2005) The mirror neuron system and imitation. Perspectives on imitation. Neurosci Soc Sci 1(1):55–76
65. Decety J (2002) Naturalizing empathy. LEncephale 28(1):9–20
66. Decety J, Jackson PL (2004) The functional architecture of human empathy. Behav Cogn Neurosci Rev 3(2):71–100
67. Gallese V, Goldman A (1998) Mirror neurons and the simulation theory of mind-reading. Trends Cogn Sci 2(12):493–501
68. Preston SD, De Waal FBM (2002) Empathy: its ultimate and proximate bases. Behavioral Brain Sci 25(1):1–20
69. Belman J, Flanagan M (2010) Designing games to foster empathy. Int J Cognitive Technol 15(1):11
70. Larson EB, Yao X (2005) Clinical empathy as emotional labor in the patient-physician relationship. JAMA 293(9):1100–1106

71. Aiko MORIWAKI, Ryoko ITO, Hiroshi FUJINO (2011) Characteristics of empathy for friendship in children with high-functioning autism spectrum disorders. Jpn J Special Educat 48(6):593–604
72. Todd AR, Galinsky AD (2014) Perspective-taking as a strategy for improving intergroup relations: evidence, mechanisms, and qualifications. Soc Pers Psychol Compass 8(7):374–387
73. McQuiggan SW, Robison JL, Phillips R, Lester JC (2008) Modeling parallel and reactive empathy in virtual agents: An inductive approach. In: Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1. pp 167–174. International Foundation for Autonomous Agents and Multiagent Systems
74. Alessio A, Ilaria M-P, Ilaria B, Aglioti Salvatore M (2009) The pain of a model in the personality of an onlooker: influence of state-reactivity and personality traits on embodied empathy for pain. Neuroimage 44(1):275–283
75. Nancy Eisenberg and Amanda Sheffield Morris (2001) The origins and social significance of empathy-related responding. a review of empathy and moral development: implications for caring and justice by ml hoffman. Soc Just Res 14(1):95–120
76. Nichols S, Stich S, Leslie A, Klein D (1996) Varieties of off-line simulation. Theor Theor Mind 24:39–74
77. Lanzetta JT, Englis BG (1989) Expectations of cooperation and competition and their effects on observers' vicarious emotional responses. J Pers Soc Psychol 56(4):543
78. Hampes WP (2010) The relation between humor styles and empathy. Eur J Psychol 6(3):34–45
79. Wang CY, Ke SY, Chuang HC, Tseng HY, Chen GD (2010) E-learning system design with humor and empathy interaction by virtual human to improve students learning. In: Proceedings of the 18th international conference on computers in education. Putrajaya, Malaysia: Asia-Pacific Society for Computers in Education.(ICCE)
80. Morin A (2007) Self-awareness and the left hemisphere: the dark side of selectively reviewing the literature. Cortex 43(8):1068–1073
81. DeGrazia D (2009) Self-awareness in animals. In: The philosophy of animal minds. Cambridge University Press, Cambridge
82. Sloman Aaron (2011) Varieties of metacognition in natural and artificial systems
83. Decety J, Jackson PL (2006) A social-neuroscience perspective on empathy. Curr Dir Psychol Sci 15(2):54–58
84. Novianto R, Williams MA (2009) The role of attention in robot self-awareness. In: RO-MAN 2009-The 18th IEEE international symposium on robot and human interactive communication. pp 1047–1053. IEEE
85. Birlo M, Tapus A (2011) The crucial role of robot self-awareness in hri. In: 2011 6th ACM/IEEE international conference on human-robot interaction (HRI). pp 115–116. IEEE
86. Michel P, Gold K, Scassellati B (2004) Motion-based robotic self-recognition. In: *2004 IEEE/RSJ international conference on intelligent robots and systems (IROS)(IEEE Cat. No. 04CH37566)*, volume 3, pp 2763–2768. IEEE
87. Bongard J, Zykov V, Lipson H (2006) Resilient machines through continuous self-modeling. Science 314(5802):1118–1121
88. Ryo S, Giorgio M, Giulio S (2010) Own body perception based on visuomotor correlation. In: 2010 IEEE/RSJ international conference on intelligent robots and systems, pp 1044–1051. IEEE
89. Steinfeld A, Fong T, Kaber D, Lewis M, Scholtz J, Schultz A, Goodrich M (2006) Common metrics for human-robot interaction. In: Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction. pp 33–40
90. Anshar M, Williams M-A (2016) Evolving synthetic pain into an adaptive self-awareness framework for robots. Biol Inspir Cognit Architect 16:8–18
91. Samsonovich AV et al (2010) Attention in the asmo cognitive architecture. Biol Inspir Cognit Architect. p 98
92. Novianto R (2014) *Flexible attention-based cognitive architecture for robots*. PhD thesis
93. Kawamura K, Dodd W, Ratanaswasd P, Gutierrez RA (2005) Development of a robot with a sense of self. In: 2005 international symposium on computational intelligence in robotics and automation. pp 211–217. IEEE
94. Dodd W, Gutierrez R (2005) The role of episodic memory and emotion in a cognitive robot. In: ROMAN 2005. IEEE International workshop on robot and human interactive communication, pp 692–697. IEEE
95. Premack D, Woodruff G (1978) Does the chimpanzee have a theory of mind? Behav Brain Sci 1(4):515–526
96. Sullivan K, Zaitchik D, Tager-Flusberg H (1994) Preschoolers can attribute second-order beliefs. Dev Psychol 30(3):395
97. Carruthers P, Smith PK (1996) Theor Theor mind. Cambridge University Press, Cambridge
98. Dvash J, Shamay-Tsoory SG (2014) Theory of mind and empathy as multidimensional constructs: neurological foundations. Top Lang Disord 34(4):282–295
99. Holopainen A, de Veld DMJ, Hoddenbach E, Begeer S (2019) Does theory of mind training enhance empathy in autism? J Autism Dev Disord 49(10):3965–3972
100. Goldstein TR, Katherine W, Winner E (2009) Actors are skilled in theory of mind but not empathy. Imagin Cogn Pers 29(2):115–133. https://doi.org/10.2190/IC.29.2.c
101. Winter K, Spengler S, Bermpohl F, Singer T, Kanske P (2017) Social cognition in aggressive offenders: Impaired empathy, but intact theory of mind. Sci Rep 7(1):1–10
102. Salazar Kämpf M, Kanske P, Kleiman A, Haberkamp A, Glombiewski J, Exner C (2021) Empathy, compassion, and theory of mind in obsessive-compulsive disorder. Psychol Psychotherapy: Theory, Res Pract
103. Ratcliffe M (2006) Folk psychology'is not folk psychology. Phenomenol Cogn Sci 5(1):31–52
104. Acharya S, Shukla S (2012) Mirror neurons: Enigma of the metaphysical modular brain. J Nat Sci Biol Med 3(2):118
105. Cynthia B, Matt B, Andrew B, Jesse G, Thomaz Andrea L (2006) Using perspective taking to learn from ambiguous demonstrations. Robot Auton Syst 54(5):385–393
106. Alvin G (2006) Simulating minds: The philosophy, psychology, and neuroscience of mindreading. Oxford University Press, Oxford
107. Perry A, Troje NF, Bentin S (2010) Exploring motor system contributions to the perception of social information: evidence from eeg activity in the mu/alpha frequency range. Soc Neurosci 5(3):272–284
108. Preston SD, Bechara A, Damasio H, Grabowski TJ, Brent Stansfield R, Mehta S, Damasio AR (2007) The neural substrates of cognitive empathy. Soc Neurosci 2(3–4):254–275
109. Riek LD, Rabinowitch TC, Chakrabarti B, Robinson P (2009) How anthropomorphism affects empathy toward robots. In: Proceedings of the 4th ACM/IEEE international conference on Human robot interaction. pp 245–246. ACM
110. Takano M, Arita T (2006) Asymmetry between even and odd levels of recursion in a theory of mind. *Proceedings of ALife X*. pp 405–411
111. Devin S, Alami R (2016) An implemented theory of mind to improve human-robot shared plans execution. In: 2016 11th ACM/IEEE international conference on human-robot interaction (HRI). pp 319–326. IEEE
112. Peters C (2006) A perceptually-based theory of mind for agent interaction initiation. Int J Humanoid Rob 3(03):321–339

113. Krach S, Hegel F, Wrede B, Sagerer G, Binkofski F, Kircher T (2008) Can machines think? interaction and perspective taking with robots investigated via fmri. PLoS ONE 3(7):e2597
114. Hiatt LM, Harrison AM, Gregory TJ (2011) Accommodating human variability in human-robot teams through theory of mind. In: Twenty-second international joint conference on artificial intelligence
115. Galinsky AD, Maddux WW, Gilin D, White JB (2008) Why it pays to get inside the head of your opponent: the differential effects of perspective taking and empathy in negotiations. Psychol Sci 19(4):378–384
116. Gerace A, Day A, Casey S, Mohr P (2013) An exploratory investigation of the process of perspective taking in interpersonal situations. J Relationsh Res 4:e6
117. Vescio TK, Sechrist GB, Paolucci MP (2003) Perspective taking and prejudice reduction: the mediational role of empathy arousal and situational attributions. Eur J Soc Psychol 33(4):455–472
118. Marvin RS, Greenberg MT, Mossler DG (1976) The early development of conceptual perspective taking: distinguishing among multiple perspectives. Child Development, pp 511–514
119. Lemaignan S, Mathieu Warnier E, Sisbot A, Clodic A, Alami R (2017) Artificial cognition for social human–robot interaction: an implementation. Artif Intell 247:45–69
120. Flavell JH (1977) The development of knowledge about visual perception. In: Nebraska symposium on motivation. University of Nebraska Press
121. Johnson M, Demiris Y (2005) Perceptual perspective taking and action recognition. Int J Adv Rob Syst 2(4):32
122. Gregory Trafton J, Cassimatis NL, Bugajska MD, Brock DP, Mintz FE, Schultz AC (2005) Enabling effective human-robot interaction using perspective-taking in robots. IEEE Trans Syst Man Cyber Part A Syst Humans 35(4):460–470
123. Kennedy WG, Bugajska MD, Marge M, Adams W, Fransen BR, Perzanowski D, Schultz AC, Trafton JG (2007) Spatial representation and reasoning for human-robot collaboration. In: AAAI, vol 7, pp 1554–1559
124. Laird JE (2001) It knows what you're going to do: adding anticipation to a quakebot. In: Proceedings of the fifth international conference on Autonomous agents, pp 385–392
125. Kennedy WG, Bugajska MD, Harrison AM, Gregory TJ (2009) Like-me simulation as an effective and cognitively plausible basis for social robotics. Int J Soc Robot 1(2):181–194
126. Fischer T, Demiris Y (2016) Markerless perspective taking for humanoid robots in unconstrained environments. In: 2016 IEEE International conference on robotics and automation (ICRA)
127. Pandey AK, Alami R (2013) Affordance graph: a framework to encode perspective taking and effort based affordances for day-to-day human–robot interaction. In: 2013 IEEE/RSJ international conference on intelligent robots and systems. pp 2180–2187. https://doi.org/10.1109/IROS.2013.6696661
128. Berlin M, Gray J, Thomaz AL, Breazeal C (2006) An organizing principle for learning in human-robot interaction. Perspective taking. AAAI 2:1444–1450
129. Gray J, Breazeal C (2014) Manipulating mental states through physical action. Int J Soc Robot 6(3):315–327
130. Daniel Batson C, Shannon E, Giovanni S (1997) Perspective taking: Imagining how another feels versus imaging how you would feel. Pers Soc Psychol Bull 23(7):751–758. https://doi.org/10.1177/0146167297237008
131. Fried I, Haggard P, He BJ, Schurger A (2017) Volition and action in the human brain: processes, pathologies, and reasons. J Neurosci 37(45):10842–10847
132. Anderson JR, Bothell D, Byrne MD, Douglass S, Lebiere C, Qin Y (2004) An integrated theory of the mind. Psychol Rev 111(4):1036
133. Gregory Trafton J, Hiatt LM, Harrison AM, Tamborello II FP, Khemlani SS, Schultz AC (2013) Act-r/e: an embodied cognitive architecture for human-robot interaction. J Human–Robot Interact 2(1):30–55

Elahe Bagheri is a research scientist at IVAI. Her research interests include social robotics, humanrobot interaction and collaboration, emotion recognition, multimodal learning, and explainable AI.