The Cognitive Bases of Anthropomorphism: From Relatedness to Empathy

The Cognitive Bases of Anthropomorphism: From Relatedness to Empathy

Gabriella Airenti

G. Airenti

Center for Cognitive Science, Department of Psychology, University of Torino, Italy

e-mail : gabriella.airenti@unito.it

Abstract Humans may react very differently with respect to mechanical devices, including robots. They can interact with them with delight or retreat in aversion or fear. According to the famous model of the uncanny valley these opposite reactions depend on the degree of familiarity that different artifacts engender in humans. The aim of my work is trying to find out the cognitive bases of familiarity, analyzing the origin of anthropomorphic projection, namely human disposition to attribute anthropomorphic features - like intentions or feelings - to artifacts. I shall discuss two concepts: relatedness and empathy, and argue that relatedness is the precondition for empathy. The fact that it is possible to attribute anthropomorphic features virtually to any object shows that resemblance is not the point. Anthropomorphism is a kind of relation that humans establish with an artifact, and in order to comprehend this phenomenon we have to focus on the relational aspect. I shall argue that what we call anthropomorphism is an extension to nonhumans of forms of interactions typical of human communication, i.e. the attribution to an artifact of the position of interlocutor in a possible dialogue. It can be shown that attributing to an artifact the position of interlocutor in a dialogue implies dealing with it as if it were endowed of the features characterizing human mind, i.e. mental states and emotions.

Keywords anthropomorphism • relatedness • empathy • communication • theory of mind

1 Introduction

Empathy is a concept that has been extensively discussed in philosophy and psychology. Many authors have tried to define empathy and to discern its different components. If the basis of empathy is affect sharing, one of the major points of discussion is the place attributed to related phenomena like simple affect contagion or perspective taking [1]. This amounts to questioning the part that cognitive processes have in empathy. Developmental psychologists consider that even if already newborns may have affective responses to others, it is only during the second year that true empathy begins to emerge, when children become able to interpret others' subjective experiences as a source of their emotions [2].

In more recent years the discovery of mirror neurons [3] has put empathy at the core of the debate in neurosciences. Brain imaging studies have shown the same activation patterns when subjects perceive emotions, for instance pain, and when they observe the same emotions in others [4, 5]. Therefore, it has been suggested that the activation of the same brain structures both during first- and third-person experience of emotions would be at the basis of empathy [6]. However, other authors have presented evidence that challenges the view that empathy might be an automatic response based on shared neural circuits [7, 8, 9]. They suggest that empathy is a much more complex phenomenon modulated by appraisal processes. De Vignemont and Singer [10] propose a contextual approach to the study of empathy, distinguishing between two types of modulation of empathy. One is the voluntary control over one's emotional responses, while the other depends on

implicit appraisal processes that might influence empathic responses. They include in the modulatory factors the intrinsic features of emotions, the relationship between the empathizer and the target, the characteristics of the empathizer and the situational context.

From these brief notes on the literature it clearly appears that empathy is a complex phenomenon with many aspects involving both affective experience and cognitive appraisal. However, this claim concerns empathy between humans. If we take the topic of this issue, i.e. artificial empathy, things become even more complicated. The first question is what does artificial empathy mean, i.e. how we may define empathy when humans and robots are concerned.

There are two different ways to see the problem of empathy with respect to the relation between humans and robots. Both imply referring to human psychology but the two approaches attribute to the study of humans a different role. In one case the hypothesis is that an artificial system in order to be able to relate to humans has to be as similar as possible to them. In this case the ambition is answering the question "What is a human?" [11, 12] with the aim to identify the fundamental human characters and possibly implement them in robots. In our case the goal would be to construct a robot that might feel empathy toward a human being, or possibly toward another robot. I do not consider that as a viable goal. The question of the possibility for an artificial machine to cause mental phenomena has been at the core of a well-known debate. Harnad [13] discussed the scope and limits of the purely symbolic models of the mind, posing "the symbol grounding problem: How can the semantic interpretation of a formal symbol system be made intrinsic to the system?". For Searle [14, 15] mental phenomena are features of the brain. He has argued that machines are programmed and programs are syntactical structures. On the contrary, minds have semantic contents. In this perspective, which he calls "biological naturalism" [16], consciousness is an emergent property of brains. "Consciousness has a first-person ontology and so cannot be reduced to anything that has third-person or objective ontology" [17].

Both Harnad's and Searle's approaches are widely controversial. To give account of all the positions and discussions that they have raised in the past years would go beyond the scope of this article. With respect to our present topic, the problem that we should find a solution to may be expressed in the following terms: how an artificial machine or a robot might acquire consciousness? This question implies a great number of further questions including the role of emotions and of subjective experience [18][1]. Actually, since Searle's Chinese room, robotics has evolved toward biologically inspired systems, based on the notions of self-organization and embodiment. This approach focuses on "the reciprocal and dynamical coupling among brain (control), body, and environment" [19]. Some researchers have argued that biologically inspired robotics should go beyond sensorimotor autonomy and try to implement emotional embodiment [20]. According to Damasio [21] emotions are to be seen as bioregulatory devices providing organisms with behaviors oriented to survival. In organisms we find different levels of life-regulation, from automatic activation of survival-oriented behaviors, to higher levels regulations. These higher levels are typical of organisms equipped with feelings. The highest level of regulation is reached by conscious organisms, defined as organisms "capable of knowing they have feelings". It is clear that such a definition of emotion as a regulatory device, in the terms proposed by Damasio would be of great interest for constructing artificial autonomous systems.

More recently, Arbib and Fellous [22] have analyzed two senses of emotion: the first refers to the organization of behavior, the second refers to emotional expression for communication and social coordination. Are both senses applicable to robots? Without completely answering this question, Arbib and Fellous argue that even if robots are mechanical devices with silicon brain operating systems, we cannot rule out the possibility that they will work as biological brains in the future. Actually, they affirm that "as we better understand biological systems, we will extract 'brain operating principles' that do not depend on the physical medium in which they are implemented". Does this mean that these principles implemented in a robot would transform it in a conscious experiencing subject? The nature of those 'principles independent from the physical medium' is rather unclear. Besides, it is hard to understand how these principles could be implemented in a robot.

Moreover, what about consciousness? For the second sense of emotion, i.e. emotional expression for social interaction, would it be sufficient that robots simulate emotions or should they be capable of

[1] I consider that Chella and Manzotti in this article give a clear presentation of the theoretical issues involved in modeling machine consciousness, even if they take a position different from mine arguing against "biological chauvinism".

knowing that they have feeelings and then of experiencing emotions?

I would draw the conclusion that, considering the present state of the research, we can only state that robots *simulate* human mental features. As far as emotions are concerned, we can imagine of implementing in a robot the basic regulatory characters of emotions. However, we have no idea how to implement the highest levels, that is experience and consciousness. Furthermore, an experience is *by definition* subjective -either you have it or you don't- and cannot be simulated. "Consciousness and the experience of consciousness are the same thing" [17]. In this sense I consider that robots cannot *be* empathic. They may at most *behave as if* they were empathic.

This necessarily confronts us with a number of questions. Firstly, what humans recognize as a display of empathy? Secondly, would robots be able to display this behavior adequately? Finally, given for granted that this is the case, would the fact that robots display empathy make humans feel more confortable when interacting with them? Actually, the fact that displaying empathy would be a desirable character for a robot is not given but should be proven.

If we take this latter perspective we study humans to understand what people expect from entities they are naturally disposed to interact with. In this case it is not given for granted that a robot should be similar to a human. Instead, we try to identify the characters that could make it a good partner. In turn, this difference in objectives can be detailed according to which characters we consider as fundamental in order to establish similarity or possible partnership, such as appearance, agency, interactive abilities, feelings, etc.

If we take De Vignemont and Singer's contextual approach about empathy [10], subscribing to the importance they attribute to modulators, it clearly appears how a category of modulators seems to be particularly relevant when humans and robots are involved, namely the relationship between the empathizer and the target. This category includes two factors, similarity and familiarity, and on these two factors and their reciprocal relations I shall focus in the following. I shall argue that comprehending how people perceive similarity and familiarity may shed light on the way humans establish relations with nonhumans and that this in turn may allow further comprehension of empathy in this particular case.

2 The uncanny valley

With respect to robot construction one of the most debated questions is how far has to go the pursuit of human-like appearance in order to inspire feelings of familiarity. The problems posed by the research of perceptive similarity between robots and humans have been enlightened by Mori's seminal work on the uncanny valley [23]. Mori maintained that similarity to humans does not necessarily produce familiarity. Paradoxically, an *almost* perfect reproduction of human features evokes more a monstrous and scary entity than a similar being with whom establishing an interaction would be desirable. The outcome is fear or repulsion instead of attraction. The problem is that robots' reproduction of human features will never be perfect. For instance, a slight variation in speed of a facial movement is sufficient to make laugh look unnatural. Thus, in a graph considering familiarity as a function of robot's appearance, as robots appear more human-like, humans' sense of familiarity increases until a point where it plunges into the uncanny valley. Mori compares this situation with theater. According to him Japanese puppets - that do not resemble to real human beings - when moving (or more precisely being moved) on stage, better succeed in producing familiarity. His conclusion is that trying to obtain a complete resemblance between robots and humans is too risky and it is more sensible to pursue a "safe" familiarity with a nonhuman-like design[2].

Actually, we can argue that Japanese puppets are a special case. They perform characters in traditional narrations well known to the audience and display very fundamental feelings like passion, greed or fear that easily evoke empathy. Then, people who go to see the Bunraku are predisposed by the setting itself to feel empathy toward characters. Narration is based on the projection of features of the real world on an imaginary one and this is why the appearance of the characters is not important, their role is just to recall something that is already in the viewer's mind.

The question is what happens in less defined situations, when we deal with entities of the real world. What may predispose humans to treat a nonhuman, object or animal, with familiarity, to feel empathy toward it? Actually, it appears that humans since a very young age may naturally attribute human features

[2] Since its original publication in 1970, Mori's work raised an intense debate. Criticisms pointed to the fact that the uncanny valley was a hypothesis not validated. Nonetheless, it proved productive and was applied to different areas of research [24]. More recently, a number of psychological studies have undertaken the empirical evaluation of the graph proposed by Mori (see for instance, [25, 26, 27]). It is impossible to make here a synthesis of the results. To summarize we can assert that what emerges is an articulated set of phenomena that changes and enriches the concept of uncanny valley without disavowing it.

to objects and animals in some situations. Analyzing these situations may help us to better describe this phenomenon. Moreover, we can use the results of experimental work made in developmental psychology to find its cognitive bases: which mental states and feelings are implied and how mental features are intertwined with perceptive features. Finally, we can discuss the conditions under which the process of anthropomorphization will be activated.

3 Anthropomorphism

Let us try to define anthropomorphism. The most common definition, taken from the dictionary, is "the attribution of human motivation, characteristics, or behavior to inanimate objects, animals or natural phenomena". Using a more precise terminology we could say that we explain nonhuman behavior as motivated by human feelings and mental states, i.e. that we interpret nonhuman behavior using human folk psychology. This kind of attribution is the sign that humans *may* include non humans in social life and it is made manifest by the fact that to nonhumans we can speak, that we may quarrel with them, scold or compliment them, etc. At the same time, it is obvious that it is not always the case. In most situations nothing of this kind happens and we deal with objects unthinking. Thus, which are the conditions eliciting the process of anthropomorphization? As noted by Mori, the core notion is familiarity and we may add that familiarity can be seen from two perspectives: as existing or pursued. For instance, we feel familiarity toward our pets, but in a completely different situation, mimicking a conversation can be a way to simulate familiarity and calm one's fear of a scary animal. This second situation is particularly interesting if the prospective partner is a robot. In fact, even in the case that the robot be perceived as potentially scary, we can imagine that the same humans that perceive it as scary may have the resources to transform fear in familiarity and to establish a relation. Thus the problem becomes which are the cognitive bases of familiarity?

In humans, the attribution of social life to objects starts with pretend play, when children are 18-month-old. Developmental psychologists consider the beginning of pretend play as a fundamental step in child development as it allows conceiving in imagination worlds alternative to the real one. The objects that children include in play and that are used to construct alternative realities are disparate, teddy bears, dolls, fiction characters like Batman or Spiderman, but also wood blocks or pebbles. They can be little or big, beautiful or ugly, have a human or animal form, be completely imaginary or replicas of humans, elaborated or very simple. It clearly appears that identifying the features that would make that an object may be considered as a good candidate for being incorporated in children's pretend play is impossible. In principle, everything can be used and transformed and any appearance and functionality can be changed into another one.

In the past the common idea was that the possibility of young children of using fantasy to transform things and giving life to them was a major difference with respect to adults. Piaget considered that young children's pervasive use of imagination was due to their inability to distinguish between physical and mental facts. According to him children are both egocentric, as they understand the world in terms of their own desires, needs, sensations, perceptions, and animist, as they attribute mental states to physical phenomena. Between these two forms of thinking there is no contradiction as they are both the result of children conceptual confusion. Children's inability to realize the subjectivity of thinking and intentionality induces them to attribute by simple analogy these mental features to any external object producing independent movements and activity [28]. Adults overcome these primitive forms of thinking, magic and superstitious, regressing to it only in particular situations, when they are taken by anxiety, fear, or strong desires [29][3].

In recent years, experimental work led developmental psychologists to see things differently. Young children are able to make the distinction between reality on the one side and pretense and fantasy on the other [31, 32, 33, 34] since the age of 3. Thus, instead of considering children' s attitude to deal with imagination as a sign of confusion between entities of different nature we can see it as the beginning of a typical human capacity of exploring different kinds of interactions using fantasy to evoke what cannot be present in reality. A broom, handled as an oar, allows to evoke a boat and to play pirates, mounted as a horse, it transforms the child into a knight. The fact that, as said before, virtually any object can be used to this aim denotes the function of the object, being the trigger of a process that is mainly mental. However, in pretend play we might consider that we are at the basis of what we have seen is theater

[3] In his later work Piaget took a different position, maintaining that children actually are able to distinguish people from physical objects because they react to the child [30].

for adults, the construction of an imaginary alternative world [35].

Let us examine another form of anthropomorphism: human relationship with animals. Actually humans have very different relations with animals. We have pets with which we interact and play. We consider them as companions and we feel duty bound to treat them following ethical rules. For instance, abandon them, inflicting them pain or any form of abuse is regarded as unethical. Other animals, that are not included in this category are treated very differently: they can be killed and some of them eaten. I cannot enter here in the very complex relationships that human entertain with animals and all the ethical and religious problems that are involved. The only thing that I want to stress here is that the features of animals, which enter in different categories differ from society to society, in the same society from a group to another, from a situation to another. For instance, in some societies eating pigs or cows is forbidden, while they are a basic food in others; eating dogs or horses is normal in some countries and unimaginable in others, etc. In our society people in general (vegans excepted) do not think that eating fish poses an ethical problem but a child who would kill her red fish would be considered cruel and surely reprimanded. At the same time, in a farm a rabbit can be the children's pet till Easter time when no one will object that it is killed for being cooked. Again, another story is how the same animal is perceived in reality and in fantasy. Real bears are big, wild and dangerous to humans but in children's stories they are little and the best imaginable companions. Running into a rat in real life would be considered an unpleasant experience for everybody, while children's literature is full of clever and amusing mice. The list of differences and incongruences about human perception of animals could be continued with many other examples. But the interesting point here is simply that it is impossible in principle to define which characteristics of an animal make it a good candidate to be treated with familiarity and care or on the contrary to be considered with indifference or repulsion.

From the two examples that I have presented, young children's pretend play and human disposition toward animals, we can draw two conclusions. A first conclusion is that human attitude to establish a relation of familiarity with an object or an animal is basic, as it appears very early in life. Secondly, it does *not* depend on particular features of that object or animal. Thus, if the source of anthropomorphism cannot be found in the object to which human features are attributed, to explain this phenomenon we have to turn to human cognition. We have to analyze the origin and development of human predisposition to attribute human mental features to nonhumans and highlight the conditions under which this attribution is actually carried out.

4 Children's theory of mind

In present cognitive psychology one of the most interesting area of research is the theory of mind. Under this name goes the study of the development of children's representations of others' minds[4].

Humans consider that their own and others' behavior is motivated by mental states, like desires and beliefs, and emotions, like love or fear. Moreover, the implementation of behavior is seen as mediated by other mental structures like perceptions, goals and plans. This is adults' fully-fledged theory of mind. Children acquire this folk vision of the mind in some years passing through a certain number of steps and precursors. The first fundamental stage in children's attribution of mental states to others can be seen in the ability to see others as intentional agents. This ability is already present in infants. By 6 months of age infants see human actions as goal directed [37] and by 9 months children expect that people have goals, they pursue them until they are reached, and they are happy or sad depending on success or failure [38].

Different authors have considered that the source of this ability has to be found in the perception of self-produced movement. We have seen that this was already Piaget's intuition [29] even if he attributed it to much older children.

Premack [39] specified the link between movement and intentionality proposing that infants are innately endowed of two properties, causality and intention. These properties allow children to distinguish the perception of movement of non self-propelled objects from the perception of movement of self-propelled objects: in the first case the state of motion is changed by another object while in the second case changes in motion are not dependent on any other object. Moreover, if the self-propelled movement leads to the perception of intentionality, under some conditions the infant can perceive one object as having the goal of affecting the other object. The importance of movement in characterizing different kinds of entities is supported by Spelke's research [40] who showed that it

[4] It is interesting to note that this area of studies had its origins in ethology. At the beginning the question was about the representations that chimpanzees had of the humans with whom they interacted [36].

is precisely movement the allows infants to identify objects since when they are 3-month-old.

Gergely and Csibra [41, 42] proposed another point of view to explain young children's ability to draw inferences about goal-directed actions. They also take into consideration the perception of movement: however, they argue in favor of a non mentalistic system of interpretation, that they call "teleological stance". By the age of 12 months children would possess a simple theory of rational action, which would allow them to interpret actions as means to attain goals in different contexts.

The discussion about the relationship between intentionality and rationality would be very interesting but goes beyond the scope of this paper. I limit to consider that, as maintained by Premack and James Premack [43], it is difficult to think that rationality does play a role in infants' attribution of intentionality, as rational thinking is a high level form of reasoning typical of adults. Moreover, taking a teleological stance can be described as perceiving an action as a means to reach a goal, and then, in my view, it is by definition a mental attitude, the attribution of an intention. This is precisely what the studies presented in [41, 42] have very brilliantly shown.

For the present purpose the interesting point is that infants perceive actions as goal-directed and that this interpretation concerns not only human actions but also actions performed by objects. In the experiments reported in [41, 42] what infants perceived as goal-directed actions were the movements of colored dots on a screen. Children saw a little yellow dot moving toward a bigger red dot. At the start along the path there was an obstacle that had to be outridden or avoided. After, in two different conditions, the obstacle had been removed. Children were puzzled when the little yellow dot did not behaved as it should be expected by an intentional (or rational, according to the authors) actor and unnecessarily avoided the most direct path to the goal.

Other experiments have shown that infants not only can see actions as goal-directed but that they also evaluate others' actions [44]. What is interesting for the present purpose is that again in this study the agents whose behavior infants evaluated were wooden blocks of different colors and shape with "googly eyes", which moved up and down a green incline. One block tried to reach the upper plateau of the incline and then another block intervened. The second block either helped the climber pushing it to the top of the hill or hindered it pushing it down to the bottom. These situations were presented to 6 and 10-month-olds. Children of both groups manifested preference for the helper with respect to the hinderer revealing the ability to make social evaluation of actions. A further study, using a similar methodology, showed that infants as young as 3 months of age evaluated others' social behavior, in particular manifesting aversion toward antisocial actors hindering others' goals [45].

What appears from the experimental work presented above is that infants are able to detect intentionality (or goal-directed action) and are disposed to evaluate actions, i.e. to consider the social value of actions. However, there is something more. In all the studies that we have mentioned the actors were objects, dots on a screen or wooden blocks. This shows that infants as soon as are able to interpret behavior as intentional and social, may in principle extend the attribution of these mental features to nonhumans.

We can then draw the conclusion that the bases of anthropomorphism are already present in the first months of life. Thus what we see in older children, who attribute feelings and mental states as desire and beliefs to objects in pretend play, has its preconditions in more basic attitudes already present in infants.

The question is now to understand how this latent attitude to attribute mental features to objects that we have seen manifesting in very specific experimental situations, actually appears in real life. We have identified a potentiality but how is this potentiality put to use? It is obvious that adults make anthropomorphic attributions only in some situations. And in fact this is the case also for children. Children may be animist, to use Piaget formulation, probably more frequently than adults, but not in any situation. For them, as for adults the attribution of mentality is submitted to conditions. For them too there is no confusion of status. As remarked by Karl Bühler [46], a child can easily throw into the fire a piece of wood that until a moment before *was* a beloved baby.

Thus, for children as for adults the attribution of human features to objects is performed in some situations under certain conditions. In the next section we shall analyze the circumstances under which this behavior actually occurs.

5 The dialogic nature of anthropomorphism

Consider the most typical cases in which we ascribe human characters to objects. As already stressed before, such an attribution is completely independent from the nature of the object. What matters are the *motivation* and the *interactive situation*.

In general our attitude is motivated by an emotion that can be negative or positive. For instance, our car stops and we are upset because we do not know how to deal with the situation or, on the contrary, we are

agreeably surprised because we feared not to arrive in time at a meeting and in fact we succeed. One way to express these emotions is to address the object as if it were a human partner. We can express our disappointment or anger insulting or cursing, but we can also pray for help or comprehension: "Now, I start one last time and you go! OK?" or "You can't do it to me today of all days". In the unexpected happy ending we can express our relief saying to our car: "I knew that I could count on you!" Obviously there is no misinterpretation possible. We perfectly know that we are dealing with an object even if a complex one. The problem is that an emotion can be expressed only in an interaction. Then, the object that in the most common situations is simply manipulated, becomes a partner in a dialogue, and this in turn means that it is attributed humanlike mental features, namely intentionality, motives, etc. We are here again in the presence of one basic aspect of human cognition, its dialogical nature, and we can try to find out its beginning in infancy [47].

We have seen that infants are able to detect others' intentionality. Infants have another basic capacity and it is the ability to interact with others in the form of a dialogue. Developmental research has shown that almost since birth infants participate with adults in interactions, which have the form of dialogues [48, 49]. Naturally, first dialogues, at least as far as the infant is concerned, are nonverbal. What is dialogic is the structure, in particular turn-taking. For instance, adults and children may engage in sequences of reciprocal imitations [50]. It is the dialogic structure that makes significant every gesture or sound that within this structure is included and that makes it possible to see them as the first manifestations of communicative behavior [51]. Note that already during the first year infants can modulate different kinds of interactions, for instance distinguishing serious from joking situations [52].

Thus, in real situations infants attribute mental features to partners in interactions. At the beginning the development of subjectivity and intersubjectivity are two relatively separate processes: children learn how to approach objects and how to interact with humans and what they can expect in these different situations. At around nine months, children that until this moment have manipulated objects and communicated with humans start to deal with both, i.e. to make objects part of interactions [53]. For instance, they may participate in a play of give and take. We can make the hypothesis that this is the precondition for attribution of mental features to objects. Once objects become part of interactions they can also be addressed as interlocutors themselves.

Meltzoff and his collaborators made a very interesting experiment in which they studied the behavior of 18-month-old infants with respect to a humanoid robot [54]. The task was to follow the robot's gaze toward an external target. It resulted that children were more likely to follow the gaze of the robot when they had seen it previously involved in a communicative interaction with an adult experimenter. They considered the robot's participation to a communicative interaction as evidence of its psychological nature. Importantly, the authors could show that neither the humanoid appearance nor autonomous movement were sufficient to this aim. The key element was the robot's involvement in a social interaction.

We can support the fact that there is a fundamental link between the attribution of mental features and communication introducing another phenomenon that starts around the end of the first year, namely *social referencing* [55, 56]. When they are in a situation of uncertainty with respect to a stimulus in the environment (be it a stranger or an unknown toy), children look at the caregiver. It is the emotion expressed by the adult that will guide the child's ensuing behavior. With their gaze children are trying to know if a relation with the object (or the person) is encouraged or not by the adult. In a simple way this behavior can be viewed as the child's need of adult's help in order to categorize a new entity: is it dangerous or good? But this attitude has other consequences. It allows or forbids a possible interaction on the basis of an accepted or refused familiarity. In developmental literature social referencing is considered as a fundamental step in the development of empathy. The recognition of the fact that external entities may arise emotions in others opens the way to being concerned by others' emotional experiences.

Following the development of the basic cognitive processes in infants we have then identified some elements that allow us to clarify the concept of familiarity with respect to anthropomorphism. Familiarity actually can be better defined as *relatedness,* the disposition to consider others as possible interlocutors in communicative interactions. This disposition manifests naturally in infants towards their caregivers. It can be extended to other humans, animals and even to objects under the "supervision" of adults.

There are two fundamental points that we can stress here. One question regards the link between communication and theory of mind. Taking an entity as a partner in a communicative interaction necessarily involves attributing it a theory of mind. This is automatic. There are no intermediate possibilities

between perceiving an object or an animal as a pure mechanism or attributing it the same mental states that we attribute to humans. The same object can be seen either as a mechanism or as an interlocutor, but when it is in the position of interlocutor it will be attributed the same folk psychology that it is attributed to humans. Humans have just one folk psychology. This means that the artifact will necessarily been attributed a goal-directed behavior but also desires and beliefs and a positive or negative attitude, being a helper or a hinderer.

There is an example coming from the early days of artificial intelligence that constitutes a clear example of this phenomenon: ELIZA, the program designed by Joseph Weizenbaum in 1966 [57]. The author's intent was to develop a program able to interact with a human user and he had the idea to give the computer the role of a psychotherapist. ELIZA had a rather simple functioning that included a limited number of possible reactions. It could ask simple questions, as for instance: "Tell me about your father", or just repeat the user's last words as a prompt: "Tell me more about..." etc. The author was amazed to discover that the users interacted with ELIZA at length and discussed very seriously their problems as if it really were a human. This was surprising because the users were students and researchers in the same laboratory of the author and then perfectly in the know of the nature of the program. The form of dialogue was sufficient to force the attribution to ELIZA of human mental states despite the awareness that ELIZA was a computer program and the simplicity of its replies.

Another interesting point that emerges from the previous considerations is the social aspect of familiarity. The phenomenon of social referencing can explain the fact that familiarity, defined as we have proposed here, in terms of the attitude to put an animal or an object in the position of interlocutor in a dialogue, is the product of social transmission. This is clear with respect to animals. As said before, any society has an important place for them but which role a particular animal has is part of cultural knowledge, and then differs from society to society and it is transmitted. It is only studying a culture in its whole that we can understand why jaguars or pigs are considered as humans' interlocutors. But it is also the case for objects. In our society cars have a central role in people's life and that can explain why so often they are addressed as if they were humans. It is just the same for computers. We can imagine that in a hunter-gatherer society a bow can be addressed as a companion. As far as robots are concerned, if they appear in movies and literature, they are still scarcely present in our everyday life. Moreover, they have a hybrid status, as they are mechanisms exhibiting an apparent autonomy of behavior.

To conclude, establishing relatedness with animals and objects and then attributing mental states to them is one human fundamental attitude but the conditions for its implementation are socially determined and transmitted.

6 Relatedness and empathy

We have argued that relatedness is a precondition for human empathy. Any possible target of empathy needs first to be considered as a possible partner in an interaction. We can now address more precisely the question of how relatedness is connected to empathy. Does empathy naturally descend from relatedness? Actually, human beings can interact with others, consider them as interlocutors and at the same time not care for them. Thus, if relatedness is a precondition of empathy, empathy needs something more. Empathy towards human beings means being affected by what of positive or negative happens to them, feel what they feel, being disposed to help or console them when they are in distress. As regards the possibility for humans of establishing a relation with a robot may we expect such a deep sense of empathy?

To get a clearer picture, we should first specify on which side empathy is supposed to be. Should robots manifest empathy towards humans or what we expect is that humans feel empathic feelings toward a robot? To answer the first question we should know how humans perceive robots manifesting human feelings. As regards the human attitude toward the robot what we could expect it may not be empathy in the full sense of term. It could be for instance, the sheer pleasure of interaction. It is important to stress that when we consider human interactions with robots we do not think of an occasional interaction as it can happen with other objects. We are actually much nearer to the case of animals. Surely, as stressed by Stephan [58], there are different types of robots and then different types of interactions that the robot is supposed to entertain with a human being. However, in all the cases the robot has to be perceived as a possible companion, assistant or guide in some activity.

It is a fact that frequently people find robots unnerving. They are so when they are too similar to humans perceptually but, as proposed by Gray and Wegner [59], they are so especially because of the mind they prompt us to see in them. Gray and Wegner in a series of experiments with adults confirmed the uncanny valley effect but they also tried to explain the reasons of this phenomenon. They consider that

humanlike appearance is so disturbing precisely because it leads to attribute to robots feelings and experience. In one of their experiments participants found a robot unnerving even when they did not see it but it was only described as having emotions. On the contrary they accepted without uneasiness the attribution of agency. The authors' conclusion is that people consider that what most characterizes human nature is the experience of emotions and that people hardly accept that this experience may be present in a robot.

Thus, we find a rather paradoxical situation. As we have seen above, humans are predisposed to anthropomorphism since infancy and they normally practice this attitude in everyday life. At the same time they do not like to see that a robot appears to them or even is simply described as endowed of human mental features. One explanation of this discrepancy reposes on a distinction between different mental attitudes. Humans would be disposed to accept robots as agents but not as experiencing beings.

I consider as the authors mentioned above, that in the phenomenon of the uncanny valley the perceptual appearance hides a problem regarding the attribution of a mind. However, I think that comparing the experience with a robot with its description turns us away from the explanation. An explanation can be found only if we distinguish these two situations, description vs. interaction.

People may act toward objects *as if* they were endowed with mental states and emotions but they do not believe that they *really* have mental states and emotions. This is true for objects in general but it is also true for robots. Obviously, normally people are unaware of the debate that developed with respect to Artificial Intelligence about what is a mind and if an artificial device can be considered as a mind [60]. However, even if they ignore the philosophical arguments used for instance by Searle [14, 15, 16, 17] to support the idea that a mind is a mind because it is implemented in human brain, they implicitly consider that a machine is a machine. Then they are uneasy when a robot that, independently of the appearance it takes, is a machine, is described as having human feelings. From a machine you can expect that it is useful and efficient in performing actions but not that it has human emotions. Then if they have to judge a description of a robot, humans consider that attributing to it the capacity of acting is reasonable, while attributing feelings it is not. This corresponds exactly to the fact that it is possible to scold a car but none would describe it seriously as having bad intentions. In fact this would be animist thinking. Why should we expect that people are animist with respect to robots?

Hence there is evidence that humans do not like that robots simulate empathy. But we can see now the other side of the problem. What would make that robots may be perceived as human-friendly? Are there features that would invite people to interact with them with pleasure and sympathy if not empathy? From what said before we can draw the hypothesis that a human will be more disposed to interact with a robot that leaves space to projection of mental states and emotions. Being too characterized as mental will make it perceived as a fake. On the contrary, humans themselves will naturally attribute mental attitudes to the robot once they are involved in interaction with it. Let us come back to ELIZA. ELIZA worked as a trigger for users' involvement with it just *because* there was nothing behind. It just proposed the structure of dialogue with an interlocutor that had neither perceptive features nor mental ones. All the contents came from the users' mind. Mentality is in humans. Humans expect that robots are agents showing goal-directed behavior. This point of view is supported also by studies based on neuroimaging. In a study was compared the brain activity while human subjects observed humans and industrial robots performing different actions [61]. In both cases the same activity was detected in the mirror neuron system. In particular, the authors showed that when the robots were engaged in meaningful human actions, their peculiar kinematics had little impact on activation. The authors' conclusion is that understanding the actions of artificial devices can take advantage of the brain mechanisms that humans have developed to understand other humans. This makes human-likeness unnecessary.

When we take into consideration more complex mental features, in particular, emotions and feelings, it is not simply that human-likeness is unnecessary. Robots that simulate them will scare people or make them think to be cheated. Instead, if we base on what we have seen already active in infants, being perceived as helpers in action will be sufficient to gain the sympathy of their human interlocutors. This will allow the users to attribute to them more complex mental states and even feelings according to the context of interaction, cultural and individual differences [62, 63].

In conclusion, if we consider empathy between humans and robots we cannot help taking into account that robots are machines and humans know it. They don't expect that robots may be empathic, as they know that they cannot experience emotions. A robot programmed to display emotions will never be as good as are humans themselves to attribute those emotions that mirror their present emotional states. Humans may interact with machines but they reserve to themselves the power to fill their mind, attributing both mental

states and emotions. Thus, we can imagine that a robot, which proves to be helpful, may be considered as an interlocutor and then attributed also emotional states. This in turn may evoke benevolent feelings on behalf of the human partner. From this, possibly even a form of empathy might arise. We can expect, for instance, that the human partner be affected by a damage that might occur to the robot.

7 Conclusions

In this paper I have tried to show how developmental psychology may contribute to enlighten what artificial empathy might be. In the literature we find many different ways to define empathy. I have considered empathy as a complex phenomenon that implies emotional and cognitive aspects. As suggested by De Vignemont and Singer [10], empathy is not the simple activation of a brain circuit but is modulated by a number of factors. I have argued that the main factor is the relation that is established between the subjects involved. This is particularly relevant when we consider empathy between humans and robots.

Humans are very precociously involved in interactions with others. This means seeing others both as agents and as bearers of benevolent or malevolent attitudes. Children since very young naturally extend these features also to nonhumans, animals and objects. Thus, we can assume that there is a natural human attitude toward anthropomorphization, i.e. attributing mental states and emotions to nonhumans and that this process naturally manifests when objects are attributed the role of interlocutors in interactions. It is on this basis that empathy can emerge. We can find an example of this phenomenon in our relations with pets. We anthropomorphize them, they become partners in interactions and empathy emerges, i.e. we care for them and we expect that they care for us.

When the interaction is with a robot we are in a different situation. If humans expect that pets have emotions, they do not expect that robots have. If they display emotions it is only simulation. In that case simulation is taken not only in the acceptation of reproduction but also in the acceptation of fake. Here we find another way to pose the problem of the uncanny valley. A robot' s display of emotions is disturbing. Thus, in this case there is no empathy on either side.

Actually, people feel uneasy when dealing with machines that are too similar to humans and they don't experience any familiarity with them. This is due to perceptual features in that perceptual features are the mirror of the mind. As maintained by [59] robots are unnerving for precisely the same reason for that they fascinate us, because they can be thought of as having a mind. Humans are delighted to treat objects as if they had a mind but they do not like that they display explicit human feelings. Emotions and feelings in human interactions are the product of the interaction itself. They change in the course of it. In interactions there is a continuous reciprocal monitoring of this process. This is not the case when robots are involved. Thus humans cannot attribute real feelings to them. What robots manifest is in no way under the control of the human interlocutor. I argue that this is why apparent human features are so disturbing.

Communication is above all reciprocity. And it is especially in the continuous reciprocal adaptation of expression of emotions, much more than in gestures and words, that reciprocity does appear. Note that we find here again something very basic for human beings. Different studies have shown that already at 12 weeks of age infants are perturbed if the mother's reaction to their communicative gestures is not adequately attuned. And so is for the mother toward them [64, 65, 66][5]. That is the real basis of empathy. In this sense empathy cannot be simulated. Paradoxically the display of human feelings hinders the process of anthropomorphization that needs space for projection. It is only in the absence of explicit reproduction of mental features that humans, once an interaction is established, may naturally attribute their own and perceive familiarity. As in the Japanese Bunraku, humans reserve to themselves the power of animating the puppets.

In this case we may expect that a secondary form of empathy may arise. Humans may at least for some moments imagine that the robot really care for them and in turn they may care for it. This form of empathy has a limited scope and it is improbable, for instance, that may evolve in behavior as it happens between

[5] In these studies mother and the infant were in two separate rooms and they interacted viewing each other in a life-sized of normal interaction, the communication was perturbed video image immediately before them. After some minutes showing to the infant mother's behaviors that occurred in a previous time and were not correlated with the present infant's behavior. While during live communication the infant behaved as in normal face-to-face interactions, in the replay phase the reaction of the infant was one of distress. In another condition it was the mother who unknowingly was presented with her infant's reaction to her previous behavior and then unrelated with her current one. Several mothers remarked that the interaction was odd and all of them changed their communication focusing more on their own experience than on the infants' one.

humans. However this limited form of empathy can be very useful in having people feel comfortable when dealing with robots.

In conclusion, in my perspective the only imaginable form of artificial empathy would be a variety of human empathy developed toward artificial artifacts.

References

1 Singer T. (2006). The neuronal basis and ontogeny of empathy and mind reading: Review of the literature and implication for future research. *Neuroscience and Biobehavioral Reviews, 30,* 855-863.

2 Thompson R. A. (1987). Empathy and emotional understanding: the early development of empathy. In N. Eisenberg, J. Strayer (Eds.), *Empathy and its development* (pp. 119-145). New York: Cambridge University Press.

3 Rizzolatti G., Fadiga L., Fogassi L., Gallese V. (1996). Premotor cortex and the recognition of motor actions. *Cogn. Brain Res. 3,* 131-141.

4 Singer T., Seymour B., O'Doherty J.P., Kaube H., Dolan R.J., Frith C.D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science 303,* 1157-1162.

5 Jackson P.L., Rainville P., Decety J. (2006). To what extent do we share the pain of others? Insight from neural bases of pain empathy. *Pain, 125,* 5-9.

6 Gallese V., Keysers C., Rizzolatti G. (2004). A unifying view of the basis of social cognition. *Trends in Cognitive Sciences, 8,* 396-403.

7 Singer T., Seymour B., O'Doherty J.P., Klaas E.S., Dolan R.J., Frith C.D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature, 439,* 466-469.

8 Lamm K., Batson C.D., Decety J. (2007). The neural substrate of human empathy: Effects of perspective-taking and cognitive appraisal. *Journal of Cognitive Neuroscience, 19,* 42-58.

9 Decety J. (2010). To what extent is the experience of empathy mediated by neural shared circuits? *Emotion Review, 2,* 204-207.

10 De Vignemont F., Singer T. (2006). The empathic brain: how, when and why? *Trends in Cognitive Sciences, 10,* 435-441.

11 Kahn PH Jr, Ishiguro, H., Friedman, B., Kanda, T., Freier N.G., Severson R.L., Miller J. (2007). What is a human? Toward psychological benchmarks in the field of human-robot interaction. *Interaction Studies*, 363-390.

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

13 Harnad S. (1990). The symbol grounding problem. *Physica D 42,* 335-346.

14 Searle J. R., 1980. Minds, brains and programs. *The Behavioral and Brain Sciences*, *3*, 417-424.

15 Searle J. R., 1990. Consciousness, explanatory inversion, and cognitive science. *Behavioral and Brain Sciences*, 13, 585-596.

16 Searle J.R. (1992). *The rediscovery of the mind.* Cambridge, Mass., MIT Press.

17 Searle J.R. (1997). *The mystery of consciousness.* New York, The New York Review of Books.

18 Chella A., Manzotti R. (2009). Machine consciousness: a manifesto for robotics. *International Journal for Machine Consciousness, 1,* 33-51.

19 Pfeifer R., Lungarella M., Iida F. (2007). Self-organization, embodiment, and biologically inspired robotics. *Science, 318,* 1088-1093.

20 Ziemke T. (2008). On the role of emotion in biological and robotic autonomy. *BioSystems 91,* 401-408.

21 Damasio A.R. (1999). *The feeling of what happens: Body, emotion and the making of consciousness.* London, Vintage.

22 Arbib M.A., Fellous J-M. (2004). Emotions: from brain to robot. *Trends in Cognitive Sciences, 8 (12),* 554-561.

23 Mori M. (1970). The uncanny valley. *Energy,* 7(4), 33-35.

24 Gee F.C., Browne W.N., Kawamura K. (2005). Uncanny valley revised. *2005 IEEE International Workshop on Robots and Human Interactive Communication,* 151-157.

25 Tinwell A., Grimshaw M., Nabi D.A., Williams A. (2011). Facial expression of emotion and perception of the Uncanny Valley in virtual characters. *Computers in Human Behavior, 27,* 741-749.

26 Saygin A.P., Chaminade T., Ishiguro H., Driver J., Frith C. (2012). The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robots actions. *Social Cognitive and Affective Neuroscience, 7,* 413-422.

27 Piwek L., McKay L.S., Pollick F.E. (2014). Empirical evaluation of the uncanny valley hypothesis fails to confirm the predicted effect of motion. *Cognition, 130,* 271-277.

28 Piaget J. (1945). La Formation du symbole chez l'enfant : imitation, jeu et rêve, image et représentation. Neuchâtel, Delachaux et Niestlé

29 Piaget J. (1926). La Représentation du monde chez l'enfant. Paris, Alcan.

30 Piaget J. 1977 (1995). *Sociological studies.* London, Routledge.

31 Wellman H.M., Estes, D. (1986). Early understanding of mental entities: A reexamination of childhood realism. *Child Development, 57,* 910-923.

32 Samuels A., Taylor M. (1994). Children's ability to distinguish fantasy events from real-life events. *British Journal of Developmental Psychology, 12,* 417-427.

33 Lillard A.S. (1994). Making sense of pretence. In C. Lewis, P. Mitchell (Eds.), *Children's early understanding of mind: Origins and development* (pp. 211-234). Hove: Erlbaum.

34 Woolley J.D. (1997). Thinking about fantasy: Are children fundamentally different thinkers and believers from adults? *Child Development*, 68, 991-1011.

35 Harris, P.L. (2000). *The work of the imagination*. Oxford, Basil Blackwell.

36 Premack D., Woodruff G. (1978). Does a chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1,* 515-526.
37 Woodward A.L. (1999). Infants' ability to distinguish between purposeful and non-purposeful behaviors. *Infant Behavior and development, 22,* 145-160.
38 Behne T., Carpenter M., Tomasello M. (2005). Unwilling versus unable: Infants' understanding of intentional action. *Developmental Psychology, 41,* 328-337.
39 Premack, D. (1990). The infant's theory of self-propelled objects. *Cognition 36*, 1-16.
40 Spelke, E. S. (1990). Principles of object perception. *Cognitive Science 14*, 29-56.
41 Gergely G., Nadasdy Z., Csibra G. Biro S. (1995). Taking the intentional stance at 12 months of age. *Cognition, 56,* 165-193.
42 Gergely, G., Csibra G. (2003). Teleological reasoning in infancy: the naive theory of rational action. *TRENDS in Cognitive Sciences 7*, 287-292.
43 Premack D., James Premack A. (1997). Motor competence as integral to attribution of goal. *Cognition, 63,* 235-242.
44 Hamlin J. K., Wynn K., Bloom P. (2007). Social evaluation by preverbal infants. *Nature 450,* 557-560.
45 Hamlin J. K., Wynn K., Bloom P. (2010). 3-month-olds show a negativity bias in their social evaluations. *Developmental Science, 13,* 923-929.
46 Bühler, K. 1930 (1999). *The Mental Development of the child*. London, Routledge.
47 Bräten S., Ed. (1998). *Intersubjective communication and emotion in early ontogeny.* Cambridge, Cambridge University Press.
48 Bateson M.C. (1979). The epigenesis of conversational interaction: A personal account of research development. In M. Bullowa (Ed.), *Before speech. The beginning of human communication*. Cambridge, Cambridge University Press, 63-67.
49 Trevarthen C. (1979). Communication and cooperation in early infancy: a description of primary intersubjectivity. In M. Bullowa (Ed.), *Before speech*. *The beginning of human communication*. Cambridge, Cambridge University Press, 321-347
50 Nadel J., Butterworth G., Ed. (1999). *Imitation in infancy.* Cambridge, Cambridge University Press.
51 Airenti G. (2010). Is a naturalistic theory of communication possible? *Cognitive Systems Research, 11*, 165-180.
52 Reddy V. (2008). *How infants know minds.* Cambridge, MA, Harvard University Press.
53 Trevarthen C., Hubley P. (1978). Secondary intersubjectivity: confidence, confiding and acts of meaning in the first year. In: A.Lock (Ed.), *Action, gesture, and symbol: the emergence of language,* London, Academic Press, 183-229.
54 Meltzoff A.N., Brooks R., Shon A.P., Rao R.P.N. (2010). "Social" robots are psychological agents for infants: A test of gaze following. *Neural Networks, 23,* 966-972.
55 Campos J.J., Stenberg C.R. (1981). Perception, appraisal and emotion: the onset of social referencing. In M.E.Lamb, L.R. Sherrod (Eds.), *Infant social cognition* (pp.273-314). Hillsdale N.J, Erlbaum.
56 Feinman S. (982). Social referencing in infancy. *Merrill-Palmer Quarterly, 28,* 445-470.
57 Weizenbaum J. (1966). ELIZA – a computer program for the study of natural language communication between man and machine. *Communications of the ACM 9*, 36-45.
58 Stephan A. Empathy for artificial agents, this issue.
59 Gray K., Wegner D.M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition, 125,* 125-130.
60 Duffy B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems, 42,* 177-190.
61 Gazzola V., Rizzolatti G., Wicker B., Keysers C. (2007). The anthropomorphic brain: The mirror neuron system responds to human and robotic actions. *NeuroImage, 35,* 1674-1684.
62 Waytz A., Cacioppo J., Epley N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science, 5,* 219-232.
63 Kaplan F. (2004). Who is afraid of the humanoid? Investigating cultural differences in the acceptance of robots. *International Journal of Humanoid Robotics, 1(3),* 1-16.
64 Murray L., Trevarthen C. (1985). Emotional regulation of interactions between two-month-olds and their mothers. In T. Field, N.Fox (Eds.), *Social perception in infants* (pp. 177-197). Norwood: Ablex.
65 Murray L., Trevarthen, C. (1986). The infant's role in mother-infant communication. *Journal of Child Language, 13*, 15-29.
66 Murray L. (1998). Contributions of experimental and clinical perturbations of mother-infant communication to the understanding of infant intersubjectivity. In S.Bråten (Ed.), *Intersubjective communication and emotion in early ontogeny* (pp. 127-143). Cambridge, Cambridge University Press.

Gabriella Airenti is associate professor of developmental psychology at the University of Torino where she has been one of the founders of the Center for Cognitive Science. She is the current president of the Italian Cognitive Science Association (AISC). She has widely published on computational and cognitive modeling of human interactions. She presently works on the development of communication and the theory of mind.