The Cognitive Correlates of Anthropomorphism

Séverin Lemaignan
Julia Fink
Pierre Dillenbourg
Computer-Human Interaction in
Learning and Instruction (CHILI)
Ecole Polytechnique Fédérale
de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
firstname.lastname@epfl.ch

Claire Braboszcz
Laboratory for Neurology &
Imaging of Cognition (LabNIC)
University of Geneva
CH-1211 Geneva, Switzerland
claire.braboszcz@unige.ch

ABSTRACT

While anthropomorphism in human-robot interaction is often discussed, it still appears to lack formal grounds. We recently proposed a first model of the *dynamics of anthropomorphism* that reflects the evolution of anthropomorphism in the human-robot interaction over time. The model also accounts for non-monotonic effects like the so-called *novelty effect*.

This contribution proposes to build upon this model to investigate the *cognitive correlates* induced by a sustained human-robot interaction and we present here our initial ideas. We propose to distinguish three cognitive phases: *pre-cognitive*, *familiarity*-based, and *adapted* anthropomorphism, and we outline how these phases relate to the phenomenological evolution of anthropomorphism over time.

1. INTRODUCTION

We recently presented a new model of anthropomorphism that focuses on the dynamics of this phenomenon [1].

Many robotics researchers tend indeed to believe that *anthropomorphism* describes a static set of human-like features of a robot (like shape, speech capabilities, facial expression). We refer to these characteristics as the *anthropomorphic design* of the robot [2]. *Anthropomorphism*, on the other hand, refers to the *social phenomenon* that emerges from the *interaction* between a robot and a user. According to Epley *et al.* [3], this includes for instance emotional states, motivations, intentions *ascribed by the user* to the robot. As such, anthropomorphism is fundamentally dynamic.

Based on a literature review which was previously published [2], a long-term field study in a natural environment [4], as well as two on-going child-robot experiments [5], we believe that *anthropomorphic effects* (*i.e.* the observable manifestations of anthropomorphism) not only evolve over time, but that they do so in non-monotonic ways. We show here how they also reflect cognitive processes experienced by the human peer when interacting with the robot.

In the following sections we first describe our model of the dynamics of anthropomorphism. We then adopt a cognitive perspective on anthropomorphism, and outline how the dynamics of anthropomorphism can be interpreted from this viewpoint.

2. DYNAMICS OF ANTHROPOMORPHISM

Figure 1 represents the phenomenological model of the long-term dynamics of anthropomorphic effects that we call the *dynamics of anthropomorphism* [1]. The model is split into three phases, depicted in different shades on the figure.

In this model, anthropomorphism is quantified by a *normalized level of anthropomorphic effects*: because anthropomorphic effects are not quantified on an absolute scale, we present them as a normalized value, that spans from a minimum (no anthropomorphic effects) to a maximum (corresponding to the novelty effect peak on Figure 1). The actual maximum value of anthropomorphic effects depends on each unique combination of human, robot and several other factors we introduce below, and thus varies. The general *shape* of the model remains however the same and depicts the evolution of anthropomorphism over time, *i.e.* the general dynamics of anthropomorphism.

The model takes into account the duration of the interaction, the nature of the interaction, as well as acquired experience and familiarization mechanisms. We also formally introduce a so-called *novelty effect* that models the first phase of human-robot interaction, during which a specific increase of anthropomorphic interactions is observed. We focus on *long-term interaction*, *i.e.* direct (non-mediated), repeated interaction with the same robot, over an extended period of time (typically longer than a week).

Initialization

During this short phase (which lasts from a couple of seconds to a couple of hours), an increase of anthropomorphic effects is observed, from the *initial capital of anthropomorphism* to a peak of anthropomorphic manifestations that corresponds to the maximum of the *novelty effect*.

The *initial capital of anthropomorphism* describes the initial potential for the robot to be anthropomorphized by the human user in a given situation. This potential depends on several factors. It has been shown, for instance, that some *people* tend to anthropomorphize more than others, that some *situations* induce anthropomorphism more than

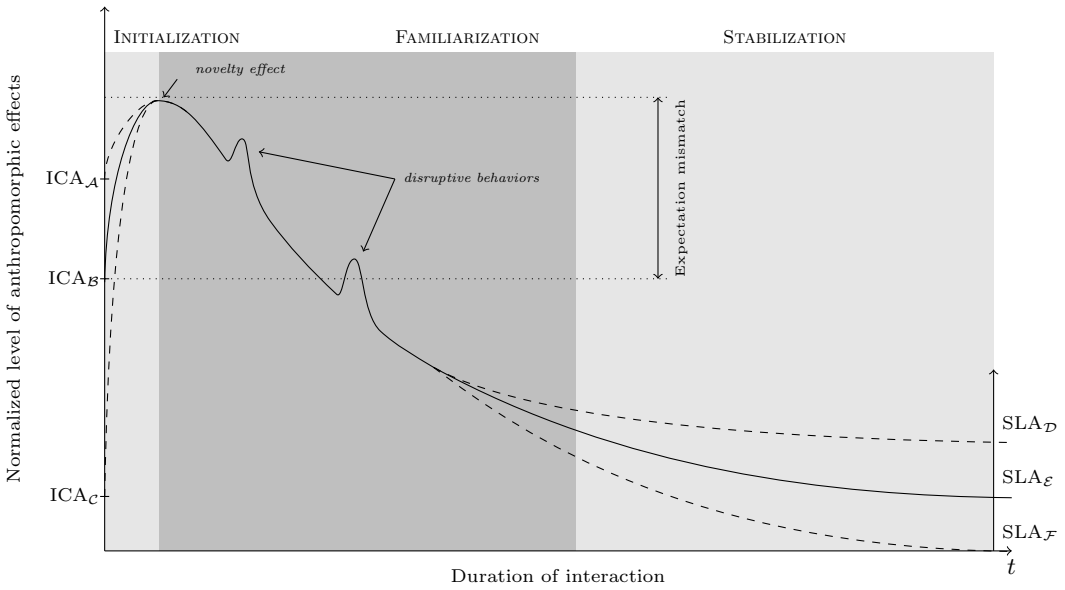

Figure 1: The dynamics of anthropomorphism. We distinguish three main phases: *initialization*, *familiarization* and *stabilization*, preceded by a *pre-interaction* phase. In the pre-interaction phase, users build an *initial capital of anthropomorphism* (ICA). Once the interaction starts, the level of anthropomorphism increases due to the *novelty effect*, and then decreases to reach a *stabilized level of anthropomorphism* (SLA). During the interaction, unpredicted behaviors of the robot (*disruptive behaviors*) may lead to local increase of the level of anthropomorphism.

others, that *children* tend to anthropomorphize more than adults, and that some *cultures* are notorious for their anthropomorphic religions and worldviews [3]. Also the shape and design of the robot play a role, and the context in which the interaction takes place. Our model of anthropomorphism takes these determinants into account and initializes the level of anthropomorphic interactions between a human and a robot to a value that we call *initial capital of anthropomorphism* (ICA). The ICA describes the first (real or imagined) contact to a robot. In this stage of pre-interaction, people form initial expectations toward the robot and imagine how they will use it / interact with it.

We build the ICA on three main factors that *a priori* determine the potential that a robot will be anthropomorphized:

1. *Human-centered factor*: The **personality** and individual traits of the human user: Psychological characteristics / determinants that influence a person's tendency to anthropomorphize artifacts [6]. Other individual traits and demographic aspects are comprised (*e.g.* age, gender, cultural background, professional background).

2. *Robot-centered factor*: The robot's **design** and how it appears to the human user. Characteristics of the robot's form, behavior, and interaction modalities (anthropomorphic design) [7].

3. *Situation-centered factor*: The real or imagined **purpose** of the robot, including the situational context in which it is used, as well as the task context and role in which the robot is used / experienced (environmental context) [8].

By taking the **purpose** of a robot into account, we suggest that the real or imagined context in which a robot is used and the interaction that this usage brings along, impacts how far the robot will be attributed human-like characteristics. We draw on findings such as presented in Joosse *et al.* [8]. The authors showed for instance that when the same robot (NAO) is used in a different task context (cleaning task *vs.* tour guide), users ascribe different "personalities" to the robot. In general, a robot which is imagined to be used in a social, entertaining or playful context leads to a higher ICA than a robot which is used for a routine or focused task (security, rescue, etc.). This idea also receives support from Goetz & Kiesler's work that revealed that people prefer a serious robot for serious tasks and a less serious robot for more playful tasks [9, 10]. Also, we suggest that the environmental context in which people experience and interact with the robot impacts the ICA. For instance, several friends interacting simultaneously with the robot might lead to increased ICAs, due to increased human-human social interactions (the robot might be perceived to be part of the social interaction, and in turn attributed human-like qualities) [11].

Familiarization

The second phase in the dynamics of anthropomorphism lasts longer (up to several days) and models the process of the human getting acquainted to the robot: by observation and interaction, the human builds a model of the robot's behavior that allows him/her to predict the robot's actions. We observe a decrease of anthropomorphic effects during this phase, that we explain by the acquired ability to predict the behavior of the robot: the initial apparent behavioral complexity vanishes, and the robot is considered

more and more as a tool.

Stabilization

The *stabilization* phase spans over a longer period of time. The level of anthropomorphic effects tends to stabilize, to reach a *stabilized level of anthropomorphism* (SLA). The SLA may be zero (no anthropomorphic effects observed anymore), but it may also remain at a higher level. The *Stabilized Level of Anthropomorphism* describes hence the long-term lasting, sustained level of anthropomorphism.

We proposed that the ICA is built on three factors: user's *personality*, robot's *design* and interaction *purpose* (or *interaction context*). The user's personality and the context of use do also influence the SLA. In particular, it appears that the user's level of acquaintance with technologies plays an important role in long-term tendency to anthropomophize [4] (people more familiar with technology understand, and hence predict, better the behavior of the robot, which in turn leads them more frequently to ultimately consider the robot as a simple tool).

The robot's design, on the other hand, plays a more subtle role, and strong initial anthropomorphic design does not mandate high SLA: lasting anthropomorphic effects have been observed on non-anthropomorphic robots (like the iRobot Roomba [4] or the military iRobot PackBot[1]), and on the contrary, anthropomorphic designs can lead to higher expectation deceptions, resulting in the robot not being used anymore.

Note that the *Initial Capital of Anthropomorphism* and the *Stabilized Level of Anthropomorphism* are generally not correlated: one individual may have high potential of anthropomorphizing (high ICA) at first sight of a good-looking humanoid robot, and get disappointed by the actual abilities of the robots, down to routine, non-anthropomorphic, interactions (low SLA), while another user with the same high ICA may, for instance, creates lasting affective bonds with the same robot, and keeps anthropomorphizing it (higher SLA).

3. COGNITIVE INTERPRETATION

We provide in this section a tentative interpretation of anthropomorphism in terms of cognitive correlates. We warmly welcome the feedback and comments from both the cognitive sciences and robotics communities to further discuss these findings.

We propose three different cognitive phases (Figure 2), which do not directly match the previously presented three phases of anthropomorphism but are still related.

Explanations for anthropomorphism

Anthropomorphism represents just one of many examples of induction whereby "people reason about an unknown stimulus based on a better-known representation of a related stimulus" [3], in this case reasoning about a non-human agent based on representation of the self or other humans.

According to Lee *et al.* [12], there are two main perspectives in explaining people's tendency to anthropomorphize. First one explains anthropomorphism from the design of the artifact. It assumes that humans directly respond to life-like or social cues that an object or system emits, without thoughtful mental processing, by simply applying stereotypes and heuristics to it. In fact, from early childhood on, humans are inherently well-trained to perceive life [6]. Schmitz [13] describes that within the visual scope of design, the outer appearance can have an important impact on the overall perception of an object. The basic assumption here is that if an artifact appears much like a human, it is likely to be treated similar to a human. If this explanation of anthropomorphism is correct, people may respond automatically to social cues emitted by a robot, and apply human-human social schemas and norms to these interactions.

The second perspective applies a human-centered, cognitive viewpoint where anthropomorphism is described through people's specific mental model they construct about how an artifact works the way it does. We then anthropomorphize because it allows us to explain things we do not understand in terms that we do understand, and what we understand best is ourselves as human beings. This is consistent with the *familiarity thesis* [14] which claims that we understand the world based upon a mental model of the world that we are most familiar with. Consequently, people tend to thoughtfully develop a mental model of agents in their environment and make inferences about it based on what is familiar to them – humans and human behavior, for instance. This point of view implicitly builds on a person's ability to attribute mental states to oneself and others (*i.e.* the availability of a *theory of mind* [15] – the link between a tendency to anthropomorphize and the engagement in the attribution of mental states to other humans has been recently demonstrated at the brain level in [16]). A theory of mind for other agents enables us to attribute intentionality to those agents [17, 18]. Previous research examined the validity of the mental model concept with various kinds of robots [13, 19]. Findings suggest that people tend to hold richer mental models of anthropomorphic robots in contrast to mechanic ones [19].

Cognitive Processes and Phases

The main underlying cognitive process in anthropomorphism is understood as perceiving and reasoning about something non-human and unfamiliar based on one's representation of the familiar and well-known concept of being human [3]. This led us to interpret the phases of anthropomorphic interactions as parallel cognitive phases (Figure 2).

The so-called *phase I* is the instinctive, pre-cognitive identification of living peers. That humans tend to anthropomorphize robots intuitively in this pre-cognitive way is supported by studies done by Rosenthal-von der Pütten *et al.* [20] who investigated the neural correlates of emotional reactions of humans towards a robot. *Empathy* is characteristic of this stage [21]. Anthropomorphism at this pre-cognitive stage might also be mediated by the human's mirror neurons system (neurons that fire both during execution of specific goal-oriented action and during the viewing of action directed toward the same goal [22, 23]) by allowing for a mapping of the robot's goal-directed actions into the human own motor repertoire [24, 25, 16].

After a longer observation period (typically including complete action sequences of the robot) or short interaction (touching, short talk like greetings), we suggest the human

[1]Rodney Brooks has reported in keynotes that occasionally soldiers would give a name to *their* PackBot and require it to be repaired instead of being replaced by another one in case of incident.

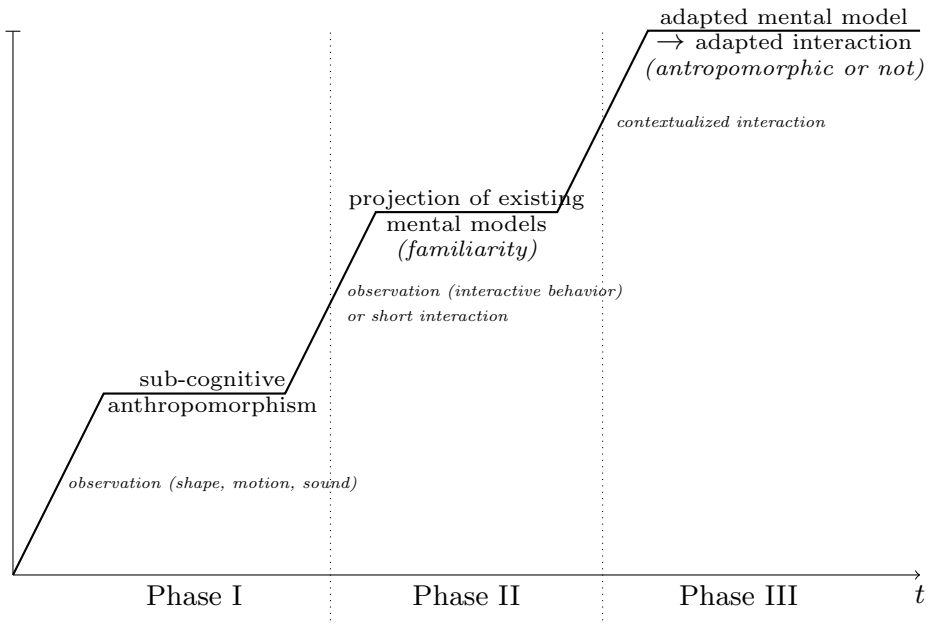

Figure 2: The three cognitive phases of anthropomorphism: Phase I is the instinctive, sub-cognitive identification of living peers. *Empathy* **is characteristic of this stage. After longer observation or short, non-contextualized interaction (typically, a lab environment), the user enters Phase II: the user projects a mental model he/she is already familiar with onto the robot. After longer** *contextualized* **interaction (typically, at home), the user enters Phase III of anthropomorphism: the user recomposes an accurate mental model of the robot, based on experience. This leads to adapted interaction modalities, that may still be anthropomorphic, or not.**

enters the cognitive *phase II*: in this phase, the human starts building a behavioral and cognitive model of the robot that would support both the observed and imagined capabilities of the robot. The *familiarity thesis* [14] supports the idea that the human first projects onto the robot mental models of similar agents he/she is already familiar with (ranging from animals to human adults, to pets and children). We hypothesize that the nature of the projected mental model, as well as how deep the human engages in this projection, might be driven by the same parameters as we mentioned for the *initial capital of anthropomorphism*.

The cognitive *phase III* occurs after a *contextualized* interaction. A *contextualized* interaction is *explicitly purposeful* (the purpose of the interaction, be it purely entertainment, is explicit and conscious to the human), and takes place in an environment that fosters a stronger cognitive (and possibly affective/social) commitment from the human in the interaction (typically, at home). During this interaction, the human iteratively restates and reshapes his/her behavioral and mental model of the robot (*How does the robot react to such and such situation/input? What does the robot know about me? About our environment? What can the robot learn?*, etc.).

This mental process depends on the human understanding of the robot's inner working, as well as his/her own tendency to anthropomorphize, but at this stage, the *perception* of the robot (its shape for instance) and its intended *purpose* play a less important role. It is mostly a human-centric process. The result of this third phase would be an iteratively adapted cognitive model of the robot.

Relation to the model of anthropomorphism

These cognitive phases overlap but do not exactly match the *Initialization*, *Familiarization* and *Stabilization* phases introduced in our model of the dynamics of anthropomorphism. In particular, cognitive phases I and II are both included in the *initialization* phase of the anthropomorphism model. Sub-cognitive anthropomorphism typically *initiates* the novelty effect by rapidly engaging the human in the interaction through an initial projected agency, whereas cognitive phase II (projection of familiar mental models) supports the novelty effect by inducing beliefs that the robot is set up with possibly complex cognitive abilities.

The cognitive phase III also overlaps with the *familiarization* phase: as the human gets used to the robot, we hypothesize one restates and adapts its cognitive model of the robot by iteratively reshaping pre-existent, familiar models until it provides a satisfying support to explain and justify the observed robot behavior.

A *stable level of anthropomorphism* is reached when the adaptation process depicted in cognitive phase III reached a stable state, *i.e.* the user's experience with the robot is correctly supported by the cognitive model he/she has built.

4. DISCUSSION

This interpretation of anthropomorphism in terms of cognitive correlates opens questions that we hope could be fruitfully discussed during the workshop. By adopting both a cognitive and a dynamic perspective on the anthropomorphic bonds that establish between a robot and a human during an interaction, some observations can be raised.

For instance, we propose that the human *adapts* iteratively to the robot by restating its cognitive model of the robot. Could our two models be conversely relied on to have the robot itself iteratively adapting its behaviour? In particular, we hypothesize that the so-called novelty effect represents a peak in anthropomorphic manifestations, followed by mostly deception-driven refinements of the cognitive behavioural model of the robot. Could we imagine that the robot pro-actively *verbalises* its own limits, in a timely manner (likely before the end of the novely effect). And how would this affect the cognitive models built by the human?

One related question we plan to elaborate on is the effects of *mutual explicitation of the mental model of the other agent*: a human and a robot interact on a given task, and after a while, we interrupt the task and ask each of the agents to explicit the mental model it has built of its partner (emotional state, beliefs, intentions, etc.). Depending on the cognitive phase the two agents are in, we may expect varying accuracy levels in the produced mental models.

Another question raised by these models relates to their transposition to human-human interaction: while concepts like *novelty effect* are not directly meaningful when analyzing human-human interaction, insights from cognitive sciences (and in particular social psychology) regarding how we, as humans, cope with unexpected behaviours from peers, or on the dynamics of mental model refinements, could bring interesting perspectives with possible applications to better manage long-term human-robot interactions.

5. CONCLUSION

This discussion on the cognitive correlates of the dynamics of anthropomorphism is still speculative, and needs to be better supported by experimental evidence.

Still, while anthropomorphism is traditionally understood as the interactions between the anthropomorphic design of a robot and the psychological determinants of the user, it appears that the duration and context of the interaction are also key factors, and that anthropomorphism needs to be understood as a dynamic phenomenon. Following this line of investigation, we propose a new formal model of anthropomorphism that accounts for these factors and introduces the concepts of *initial capital* and *stabilized level of anthropomorphism* as compound factors to characterize the profile of a given anthropomorphic interaction.

We discuss more specifically the cognitive correlates of anthropomorphism, and propose to identify three cognitive phases corresponding to successive refinements of the mental models of the robot that the user builds during the interaction. We show how these phases relate to observable anthropomorphic effects, and how they evolve over time.

While subject to discussion and further extensions, we hope that this contribution consolidates the scientific grounds of anthropomorphism, and provides support for a better understanding of long-term acceptance of robots in human environments.

Acknowledgements

This research was supported by the Swiss National Science Foundation through the National Centre of Competence in Research Robotics.

6. REFERENCES

[1] S. Lemaignan, J. Fink, and P. Dillenbourg, "The dynamics of anthropomorphism in robotics," in *Proceedings of the 2014 Human-Robot Interaction Conference*, 2014, to appear.

[2] J. Fink, "Anthropomorphism and human likeness in the design of robots and human-robot interaction," in *Social Robotics*, ser. Lecture Notes in Computer Science, S. S. Ge, O. Khatib, J.-J. Cabibihan, R. Simmons, and M.-A. Williams, Eds. Springer Berlin Heidelberg, Jan. 2012, no. 7621, pp. 199–208.

[3] N. Epley, A. Waytz, S. Akalis, and J. T. Cacioppo, "When we need a human: Motivational determinants of anthropomorphism," *Social Cognition*, vol. 26, no. 2, pp. 143–155, Apr. 2008.

[4] J. Fink, V. Bauwens, F. Kaplan, and P. Dillenbourg, "Living with a vacuum cleaning robot," *International Journal of Social Robotics*, pp. 1–20, 2013.

[5] J. Fink, P. Rétornaz, F. Vaussard, F. Wille, K. Franinović, A. Berthoud, S. Lemaignan, P. Dillenbourg, and F. Mondada, "Which robot behavior can motivate children to tidy up their toys? design and evaluation of "ranger"," in *Proceedings of the 2014 Human-Robot Interaction Conference*, 2014, to appear.

[6] N. Epley, A. Waytz, and J. T. Cacioppo, "On seeing human: A three-factor theory of anthropomorphism." *Psychological Review*, vol. 114, pp. 864–886, 2007.

[7] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and Autonomous Systems*, vol. 42, no. 3–4, pp. 143–166, Mar. 2003.

[8] M. Joosse, M. Lohse, J. G. Pérez, and V. Evers, "What you do is who you are: The role of task context in perceived social robot personality," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, Karlsruhe, Germany, 2013, pp. 2126 –2131.

[9] J. Goetz and S. Kiesler, "Cooperation with a robotic assistant," in *CHI '02 extended abstracts on Human factors in computing systems*, ser. CHI EA '02. New York, NY, USA: ACM, 2002, p. 578–579.

[10] J. Goetz, S. Kiesler, and A. Powers, "Matching robot appearance and behavior to tasks to improve human-robot cooperation," in *The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003.* IEEE, Nov. 2003, pp. 55– 60.

[11] P. Baxter, J. de Greeff, and T. Belpaeme, "Do children behave differently with a social robot if with peers?" in *International Conference on Social Robotics (ICSR 2013)*, October 2013.

[12] S.-l. Lee, I. Y.-m. Lau, S. Kiesler, and C.-Y. Chiu, "Human mental models of humanoid robots," in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation, 2005. ICRA 2005.* IEEE, Apr. 2005, pp. 2767– 2772.

[13] M. Schmitz, "Concepts for life-like interactive objects," in *Proceedings of the fifth international conference on Tangible, embedded, and embodied interaction*, ser. TEI '11. New York, NY, USA: ACM, 2011, p. 157–164.

[14] F. Hegel, S. Krach, T. Kircher, B. Wrede, and G. Sagerer, "Understanding social robots: A user

study on anthropomorphism," in *Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE International Symposium on*, Aug. 2008, pp. 574 –579.

[15] D. Premack and G. Woodruff, "Does the chimpanzee have a theory of mind," *Behavioral and Brain sciences*, vol. 1, no. 4, pp. 515–526, 1978.

[16] H. Cullen, R. Kanai, B. Bahrami, and G. Rees, "Individual differences in anthropomorphic attributions and human brain structure," *Social Cognitive and Affective Neuroscience*, 2013.

[17] A. M. Leslie, "Pretense and representation: The origins of "theory of mind."," *Psychological Review*, vol. 94, no. 4, pp. 412–426, 1987.

[18] H. Admoni and B. Scassellati, "A multi-category theory of intention," in *Proceedings of COGSCI 2012*, ser. COGSCI 2012, Sapporo, Japan, 2012, pp. 1266–1271.

[19] S. Kiesler and J. Goetz, "Mental models of robotic assistants," in *CHI '02 extended abstracts on Human factors in computing systems*, ser. CHI EA '02. Minneapolis, MN, USA: ACM, 2002, p. 576–577, ACM ID: 506491.

[20] A. M. Rosenthal-von der Pütten, N. C. Krämer, L. Hoffmann, S. Sobieraj, and S. C. Eimler, "An experimental study on emotional reactions towards a robot," *International Journal of Social Robotics*, vol. 5, no. 1, pp. 17–34, Jan. 2013.

[21] A. M. Rosenthal-von der Pütten, F. P. Schulte, S. C. Eimler, L. Hoffmann, S. Sobieraj, S. Maderwald, N. C. Krämer, and M. Brand, "Neural correlates of empathy towards robots," in *Proceedings of the 8th ACM/IEEE International Conference on Human-robot Interaction*, ser. HRI '13. Piscataway, NJ, USA: IEEE Press, 2013, pp. 215–216. [Online]. Available: http://dl.acm.org/citation.cfm?id=2447556.2447644

[22] G. Rizzolatti, L. Fadiga, V. Gallese, and L. Fogassi, "Premotor cortex and the recognition of motor actions," *Cognitive brain research*, vol. 3, no. 2, pp. 131–141, 1996.

[23] J. M. Kilner, A. Neal, N. Weiskopf, K. J. Friston, and C. D. Frith, "Evidence of mirror neurons in human inferior frontal gyrus," *The Journal of Neuroscience*, vol. 29, no. 32, pp. 10 153–10 159, 2009.

[24] V. Gallese and A. Goldman, "Mirror neurons and the simulation theory of mind-reading," *Trends in cognitive sciences*, vol. 2, no. 12, pp. 493–501, 1998.

[25] D. M. Wolpert, K. Doya, and M. Kawato, "A unifying computational framework for motor control and social interaction," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 358, no. 1431, pp. 593–602, 2003.