The Dynamics of Anthropomorphism in Robotics

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

***Abstract*—While anthropomorphism in robotics is a commonly discussed trait of HRI, it paradoxically lacks formal grounds. Supported by an extensive literature review, a long-term field study and two other on-going experiments, this report gives a first overview of a formal model of anthromorphism that we are currently building. Going beyond the traditional perception of anthropomorphism as a static feature of a system, we propose to understand anthropomorphism as a dynamic, non-monotonic and context-dependent process, which evolves over time, and has deep cognitive implications.**

I. The Dynamics of Anthropomorphism

Many robotics researchers tend to believe that *anthropomorphism* describes a 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 [1]. *Anthropomorphism*, on the other hand, refers to the *social phenomenon* that emerges from the *interaction* between a robot and an user. According to Epley *et al.* [2], this includes for instance emotional states, motivations, intentions *ascribed by the user* to the robot.

So far, the HRI community has not much investigated how anthropomorphism in human-robot interactions evolves over time (during the process of *adopting* a robot, for instance). Anthropomorphism is traditionally perceived in robotics as a static feature that once observed during a short-term interaction reflects a sustaining social effect. Based on an literature review previously published [1], a long-term field study in a natural environment [3], as well as two on-going child-robot experiments, we believe that anthropomorphic effects do evolve over time in non-monotonic ways, and play a central role in sustaining engagement in human-robot interaction.

We propose to formalize these dynamic effects into a model that we call the *dynamics of anthromorphism*. We hope that it may support further discussions and reflections on how anthromorphism impacts human-robot interaction on the long run, and also foster research on the affective bonds induced by anthorpomorphic projections on robots.

As such, we propose to represent how the level of anthropomorphic effects (*i.e.* observable manifestations of anthropomorphism) evolves over a long-term human-robot interaction (Figure 1). By long-term interaction, we mean direct (non-mediated), repeated interaction with the same robot, over an extended period of time (typically longer than a week).

Three phases

We distinguish three main phases, depicted in different shades in Figure 1.

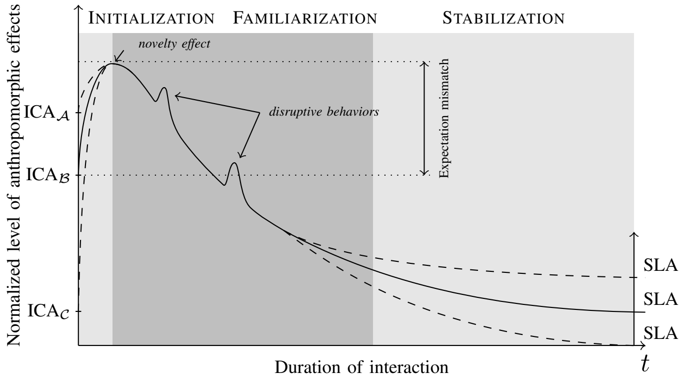

Fig. 1. Dynamics of anthropomorphism. We distinguish three main phases: *initialization*, *familiarization* and *stabilization*, preceded by a *pre-interaction* phase. In the pre-interaction phase, users build an *initial capital of anthropomorphism* (ICA). Once the interaction starts, the level of anthropomorphism increases due to the *novelty effect* [4], 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.

First, the *initialization* phase. During this short phase (from a couple of seconds to a couple of hours), we observe an increased level of anthropomorphism, from an *initial capital of anthropomorphism* (*ICA*, a value that measures initial expectations toward the robot, computed from three factors: the personality of the human, the design of the robot and the context/purpose of the interaction. The variability of this value is pictured in Figure 1) to a peak of anthropomorphic manifestations that corresponds to the maximum of the *novelty effect*.

The second phase, *familiarization*, 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.

The last phase is the *stabilization* phase. The level of anthropomorphic effects tends to stabilize over a longer time, 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. Like the ICA, the SLA is a multi-factor value that depends on the human, the robot and the interaction context.

	Unplanned by the robot	Planned by the robot
Perceived as non-intentional	case I	case IIIa
Perceived as intentional	case IIIb	case II

Fig. 2. Behaviors of the robot that are unexpected by the user may be intentional (the robot has planned the behavior) or not (typically, a failure: misdetection, bug,...). Independently of that, the behavior may be *perceived* by the user as intentional or not.

II. Elements of Interpretation

To understand and possibly better manage the evolution of human-robot interactions over time, we briefly describe in this section preliminary behavioral observations and cognitive interpretations that are related to our model.

We are currently conducting more experiments to further support and extend these findings.

a) Role of disruptive behaviors: By *disruptive behaviors*, we mean any behavior exhibited by the robot that is unexpected for the user: for instance, a robot may usually follow always the same route to go from one place to another, but suddenly change it. As long as this reason is not immediately *intelligible* to the user, the behavior counts as *disruptive*.

Our model represents *disruptive behaviors* as local increases of anthropomorphic effects in the familiarization phase (and to a lower extent, in the stabilization phase): because such behaviors are unexpected, they impact the rationality that the user ascribes to the robot, and a human observer may interpret them as the result of a richer deliberative process, which in turn leads to the supposition of complex cognitive skills in the robot. This is however to be modulated depending on the real and perceived reasons for the unexpected behavior. It may increase if the user perceives (rightfully or not) a form of intentionality, as well as decrease if the unexpected behavior is perceived as a failure. Figure 2 summarizes the possible situations.

b) Cognitive interpretations: The 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 [2]. This led us to interpret the phases of anthropomorphic interactions as parallel cognitive phases.

The so-called *phase I* is the instinctive, pre-cognitive identification of living peers. This is supported by studies done by Rosenthal-von der Pütten *et al.* [5] who investigated the neural correlates of emotional reactions of humans towards a robot. *Empathy* is characteristic of this stage.

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 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* [6] would support 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).

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 heavily 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.

These cognitive phases overlap but do not exactly match the *Initialization*, *Familiarization* and *Stabilization* phases introduced in our model of the dynamics of anthromorphism, and we are currently investigating the relations between both.

III. Conclusion

Anthropomorphism is traditionally understood as the interactions between the anthropomorphic design of a robot and the psychological determinants of the user. We have found out that the duration and context of the interaction is a third factor that plays a key role. In this preliminary report, we sketch a new formal model of anthropomorphism that accounts for these three factors and also explicits the dynamics of anthromorphism. We introduce the concepts of *initial capital* and *stabilized level of anthropomorphism* as compound factors to characterize the profile of a given anthropomorphic interaction.

While not definitive, we hope that this contribution may ultimately consolidate the scientific grounds of anthropomorphism, and provides support for further research on long-term acceptance of robots in human environments.

References

[1] 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.

[2] 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.

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

[4] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro, "Interactive robots as social partners and peer tutors for children: a field trial," *Hum.-Comput. Interact.*, vol. 19, no. 1, p. 61–84, Jun. 2004.

[5] 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.

[6] 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.