Would a Robot Trust You? Developmental Robotics Model of Trust and Theory of Mind

Samuele Vinanzi*, Massimiliano Patacchiola*, Antonio Chella† and Angelo Cangelosi*
*Centre for Robotics and Neural Systems, Plymouth University, Plymouth, United Kingdom
†RoboticsLab, Università degli Studi di Palermo, Palermo, Italy

Correspondence: Samuele Vinanzi *samuele.vinanzi@plymouth.ac.uk*

***Abstract*—Trust is a critical issue in human-robot interaction: as robotic systems gain complexity, it becomes crucial for them to be able to blend in our society by maximizing their acceptability and reliability. Various studies have examined how trust is attributed by people to robots, but less have investigated the opposite scenario, where a robot is the trustor and a human is the trustee. The ability for an agent to evaluate the trustworthiness of its sources of information is particularly useful in joint task situations where people and robots must collaborate to reach shared goals. We propose an artificial cognitive architecture based on the developmental robotics paradigm that can estimate the reliability of its human interactors for the purpose of decision making. This is accomplished using Theory of Mind (ToM), the psychological ability to assign to others beliefs and intentions that can differ from one's owns. Our work is focused on an humanoid robot cognitive architecture that integrates a probabilistic ToM and trust model supported by an episodic memory system. We tested our architecture on an established developmental psychological experiment, achieving the same results obtained by children, thus demonstrating a new method to enhance the quality of human and robot collaborations.**

***Keywords*—trust, theory of mind, episodic memory, cognitive robotics, developmental robotics, human-robot interaction**

Trust is a central component of social interactions between both humans and robots. It can be defined as the willingness of a party (the trustor) to rely on the actions of another party (the trustee) with the former having no control on the latter [1]. The fundamental role of trust evaluation is to ensure successful relationships, especially during shared goal interactions where all the parties must cooperate in a joint task to reach a common objective. The development of trust during childhood is still under debate, but one of the most interesting theories is the "trust vs mistrust" stage by Erikson [2], which states that the propensity to trust is proportionate to the quality of cares received during infancy. A psychological trait that relates to the mastery of one's self trustfulness is Theory of Mind (ToM), the ability to attribute mental states to others, as for example beliefs and intentions, that can differ from one's owns. Vanderbilt et al. [3] have demonstrated that children are not good at identifying misleading sources of information until their fifth year of age, when their ToM fully develops. Following these psychological results, we designed an artificial cognitive architecture for a Softbank Pepper humanoid robot that uses a probabilistic approach first theoretically proposed by Patacchiola et al. [4] to model trust and ToM in order to estimate the reliability of its informants. In particular, inference is computed on the probability distribution of a Bayesian network's nodes. We tested this architecture replicating Vanderbilt's experiment [3], which consists in a sticker finding game where the child, or in our case the robot, must face and learn to distinguish helpers and trickers. Our system is able to generate a belief network for each user and to perform decision making and belief estimation. In addition, an episodic memory module makes the robot able to build a personal character that depends on how it has been treated in the past, thus making it more or less keen to trust someone it never met. The results we obtained are in line with the original experiments, thus confirming that our architecture correctly modeled trust and ToM mechanisms in a humanoid robot. In the future, we plan to use this model in a wider scenario where trust estimation and intention reading will generate and modulate collaborative behavior between humans and robots.

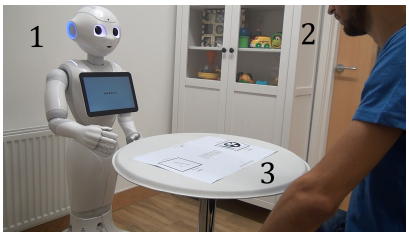

Fig. 1. Experimental setup. A Pepper robot (1) and an informant (2) face each other in front of a table where a sticker can be moved between two positions (3).

Acknowledgment

This material is based upon work supported by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF under Award No. FA9550-15-1-0025.

References

[1] Roger C Mayer, James H Davis, and F David Schoorman. An integrative model of organizational trust. *Academy of management review*, 20(3):709–734, 1995.

[2] Erik H Erikson. *Childhood and Society*. W. W. Norton & Company, 1993.

[3] Kimberly E Vanderbilt, David Liu, and Gail D Heyman. The development of distrust. *Child development*, 82(5):1372–1380, 2011.

[4] Massimiliano Patacchiola and Angelo Cangelosi. A developmental bayesian model of trust in artificial cognitive systems. In *Development and Learning and Epigenetic Robotics (ICDL-EpiRob), 2016 Joint IEEE International Conference on*, pages 117–123. IEEE, 2016.