A Global Workspace Theory Model for Trust Estimation in Human-Robot Interaction

Francesco Lanza[1 [0000−0003−4382−6366]], Samuele Vinanzi[2 [0000−0003−0241−9983]], Valeria Seidita[1,3 [0000−0002−0601−6914]], Angelo Cangelosi[2,3 [0000−0002−4709−2243]], and Antonio Chella[1,3 [0000−0002−8625−708X]]

[1] Dipartimento di Ingegneria, Università degli Studi di Palermo, Palermo, Italy
name.surname@unipa.it
[2] School of Computer Science, The University of Manchester, Manchester, United Kingdom
name.surname@manchester.ac.uk
[3] ICAR-CNR, Istituto di Calcolo e Reti ad Alte Prestazioni,Palermo, Italy

Abstract. Successful and genuine social connections between humans are based on trust, even more when the people involved have to collaborate to reach a shared goal. With the advent of new findings and technologies in the field of robotics, it appears that this same key factor that regulates relationships between humans also applies with the same importance to human-robot interactions (HRI). Previous studies have proven the usefulness of a robot able to estimate the trustworthiness of its human collaborators and in this position paper we discuss a method to extend an existing state-of-the-art trust model with considerations based on social cues such as emotions. The proposed model follows the Global Workspace Theory (GWT) principles to build a novel system able to combine multiple specialised expert systems to determine whether the partner can be considered trustworthy or not. Positive results would demonstrate the usefulness of using constructive biases to enhance the teaming skills of social robots.

Keywords: Trust · Emotions · Global Workspace Theory · Theory of Mind · Human-Robot Interaction · Cognitive Robotics

1 Introduction

Most of our lives depend on trust: our ability to manage it, granting or denying it whenever appropriate, is critical for our safety and well-being during our everyday lives. Misplaced trust can have catastrophic effects over our physical, emotional or economic welfare. Being such an important factor for humans, it is natural to think that it would also benefit the social robots we are hoping to involve in our future relationships. In particular, regarding shared goal scenarios where a human and a robot need to interact and collaborate to achieve a common objective, it is important that both the involved agents are able to estimate the trustworthiness of each other so to adopt the best decisions that

would ensure the successful completion of the task. Castelfranchi and Falcone [6] argue about the importance of trust in the context of society: we, as humans, use it to define our behaviour, to guide our decisions and in general to build successful relationships. The use of trust dynamics doesn't have to be limited to human interactions but can be extended, with all its virtues, to relationships with artificial agents. Because of this, robots could improve their performance by being able to perform decisions based on trust.

In this paper we focus on the trust that is assigned from a robot to a human, that means that the former will be the trustor and the latter will be the trustee. We consider an existing computational model of trust and Theory of Mind (ToM) proposed by Vinanzi et al. [24] and we propose an extension which refines the trust estimation procedure by taking into account several social cues, as for example emotions and gaze direction. Our objective is to overcome the two main limitations of this current model, which are: the unimodality of the perceptual information used and its sole reliance on personal experience. The architecture we propose uses machine learning techniques to gather the previously mentioned social features from an HRI scenario and uses the GWT principles to determine whether the partner should be trusted or not during the joint action engagement.

This proposal paper is structured as follows. Section 2 provides an overview of the current literature into which this work taps in. Section 3 describes our proposed design of a cognitive architecture that makes use of GWT to extend Vinanzi's model [24], Finally, Section 4 provides a discussion on the expected results, the future works and places this piece of research in a wider, incremental project.

2 Theoretical Background

Trust has been studied extensively by a wide range of researchers spacing across different fields including but not limited to psychology, computer science, business and law [10]. The interdisciplinary interest towards this topic is indicative of its importance on many levels of human society. Mayer [18] defines trust as the willingness of a party (the trustor) to rely on the actions of another party (the trustee) with the former having no control over the latter. In other words, in a trust-based scenario the trustor accepts a vulnerability in hope of a positive but uncontrolled outcome. This is particularly true regarding teaming and cooperation [17].

It has been demonstrated that the ability to correctly judge the trustworthiness of others is strongly correlated to a cognitive skill known as Theory of Mind (ToM): the ability to attribute mental states to others (for instance intentions, beliefs and desires), that may differ from one's owns [21]. Vanderbilt [23] proved that this skill gradually develops through childhood and matures fully around the fifth year of age by designing an experiment involving a sticker finding game: some children of different ages needed to locate a sticker hidden in one of two locations by relying on the suggestion provided by an adult, who could either be a helper or a tricker. The children had time to familiarise with their informant

and subsequently decide whether to trust them or not based on the experienced behaviour.

Whereas trust is fundamental in human relationships, it's also a key factor in HRI: a lack of trust in the robot's skills or, vice versa, an overestimation of it's capabilities can both negatively impact the cooperation's performance. Just as human relationships are never unidirectional, even those with robots should be: previous works have addressed the importance of models of trust in cognitive robotics. We will focus in particular on an integrated model of trust and Theory of Mind (ToM) for social robots developed by Vinanzi et al. [24], with the aim of extending its capabilities beyond its original scope.

This model is based on an established framework known as developmental robotics. Cangelosi [5] defines it as "the approach to the design of behavioral and cognitive capabilities in artificial agents that takes direct inspiration from the developmental principles and mechanisms observed in the natural cognitive systems of children". Following this approach, the model is based on the same psychology experiment performed by Vanderbilt [23] to test the maturity of ToM and is able to learn how to distinguish a trustworthy and untrustworthy informant. To do so, the cognitive architecture makes use of an unified model of trust and ToM originally developed by Patacchiola [19] based on Bayesian Belief Networks (BBN) and extends it with episodic memory, which is a subcategory of the long-term declarative memory that stores memories about temporally dated episodes or events and temporal-spatial relations among them [22]. Thanks to this extension, the robot is able to interact with informants with who it has never familiarised, using its developmental experience to decide whether to instinctively trust them or not.

This model is based on the personal experience of the agent, which is composed by its current perception and the history of its interactions with several other agents. In other words, it uses a unimodal perception information to produce the trust evaluation. Our plan is to further extend this model to take into account other social cues from the human partner with the final objective to refine even more the trust estimation task. We think that by transforming the process to take into account multimodal sensory information we will be able to achieve a much higher performance.

According to Cho [10], given the multidisciplinary nature of trust, different kind of factors affect its evaluation. Of these, we are going to consider the ones which influence cooperation and collaboration in a human-robot teaming interaction. For example, Ekman [11] describes how emotions and expressions represent the key to read human intentions. Facial expressions provide behavioural and situational information in trust contexts [4]. He has developed an "atlas of emotions" to associate emotional feelings to emotional expressions [20]. Even if emotions and expressions can be considered good sources of information to predict the trustworthiness of a person, Ekman suggests to integrate other factors, which he calls "macro-expressions" and includes: symbolic gestures, tone of voice, demographic data and content of speech [12]. He then presents some ex-

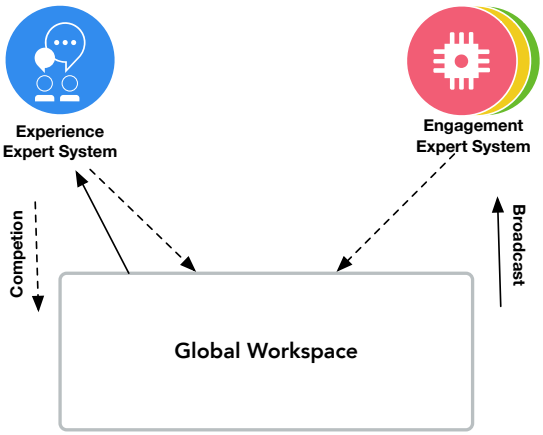

Fig. 1. The GWT model to predict trust in HRI. It is composed by two expert systems, both of which use different features to estimate the informant's trustworthiness. The Experience Expert System (ExES) uses previous experiences of interactions, while the Engagement Expert System (EgES) makes use of human-like features detected on the partner during the interaction. The final decision of the system will be moderated by the Global Workspace (GW).

periments which demonstrate that it is possible to use these features in a human interaction context to modify the perceived trustworthiness of someone.

The cognitive model is extended using Trust Theory [6], based mainly on delegation and adoption concepts [7, 14, 15], and GWT. The latter is a cognitive model proposed by Baars [2, 3] which is described metaphorically as a theatre where several actors (the working memory, ensemble of expert systems) compete between them to earn the "spotlight of selective attention" on stage (the consciousness), while most of the background work remains invisible and behind the stages (the unconscious) [1].

3 Proposed Model

The GWT model depends on the interaction between several specialised nodes of a network, which are the expert systems that compete for the spotlight of the artificial consciousness. Figure 1 shows the proposed architecture based on GWT. The cognitive model hosts two expert systems: the *Experience Expert System* ExES and the *Engagement Expert System* EgES, which are described in detail in the following Sections and are both capable of providing trust estimations. The GW component is in charge of deciding whether to assign the spotlight to one system or the other. This architecture is modular, meaning that the expert systems can be changed in typology and number based on specific needs.

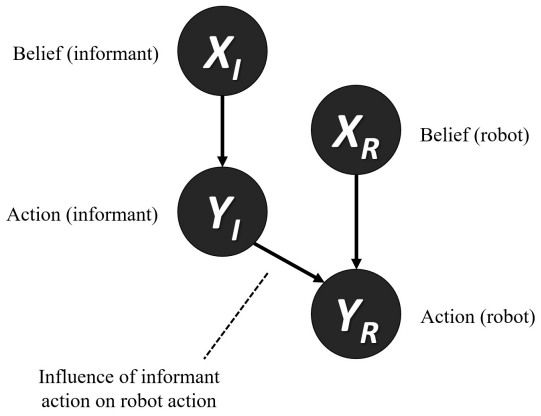

Fig. 2. The BBN that models the relation between the robot and an informant within the ExES. The agent generates a separate network for each informant, with the same structure but different probability distribution.

3.1 Experience Expert System

The ExES [24] is a developmental robotics cognitive architecture that enables a robot to perform trust evaluations based on personal experience. It's core component is a BBN that unifies trust and ToM considerations for the sticker finding game, as described in Figure 2. Following the original ToM experiment [23], it performs two main functions:

1. **Familiarisation**: the robot experiences some interactions with the informants, observes the outcome of their suggestions learning who should be considered a helper or a tricker and trains a BBN for each of them;
2. **Decision making**: the robot will use probabilistic inference on the BBNs to decide whether to follow or reject the suggestion for each particular informant. See Figure 3 for a visual demonstration applied to the Vanderbilt experiment [23];

In addition to this, and following the psychological principles of trust development defined by Erikson [13], the robot is able to use its Episodic Memory to collect all of its past experiences and decide how to act towards a novel informant it has never familiarised with. The history of interactions shape the robots personal character development, which will determine its attitude to instinctively trust or distrust a stranger.

3.2 Engagement Expert System

The EgES is a cognitive model under development whose purpose is to estimate the trustworthiness of the informant using social cues. According to Ekman [11],

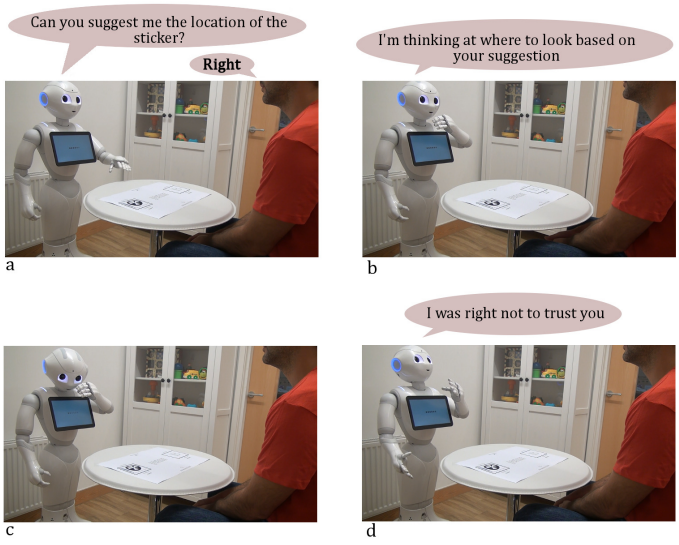

Fig. 3. Decision making phase with a tricker informant. (a) The robot asks for a suggestion on the location of the sticker and receives a misleading suggestion from the informant. (b) The robot performs inference. (c) The agent decides not to trust. (d) The robot finds the sticker and gives feedback to the informant. Adapted from [24].

emotions and micro-expressions are used as subtle hints to understand whether someone wants to trick us. Using this principle, EgES feeds the robot's RGB camera images and auditory perceptions to a stack of machine learning algorithms that classify the current perception and extract a set of features for an artificial neural network (ANN). The latter is used to predict the appropriate level of trust to assign to the informant under analysis. Figure 4 shows a schematic representation of this process. The computed features are the following:

- **Emotions**: the emotional state of the informant. This system classifies emotions as sad, angry, happy, or neutral;
- **Vocal Emotions**: the emotion expresses by the informant's voice signal, without considering its content;
- **Facial Action Units**: facial movements and micro-expressions;
- **Gaze**: the direction of the informant's gaze, to determine if the informant is looking directly at the robot or not;
- **Gender**: different genders show differences in their social cues, this feature takes this into account;
- **Age**: as above, social cues differ with age;
- **Context**: the state of the environment, such as calm or agitated.

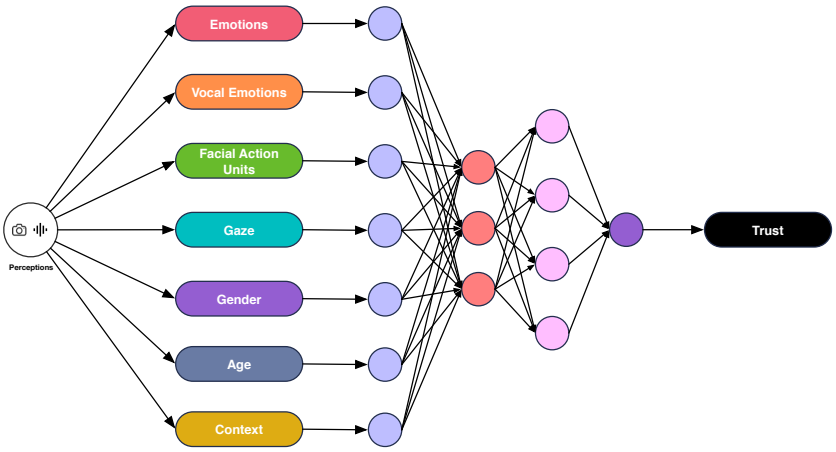

Fig. 4. The EgES architecture. Perceptions are processed by a stack of machine learning algorithms which will compute a set of features. The latter are processed by an ANN to determine the final level of trust

3.3 Competition of Expert Systems

At this point of the paper, it is clear that our cognitive model based on GWT is built upon two expert systems which are able to independently estimate the trustworthiness of another agent engaged in joint action. These two systems will compete to earn the spotlight of selective attention by the GW, which represents the artificial consciousness of the system. The latter is a mathematical model whose purpose is to select the results of either the ExES or the EgES to apply to a specific situation. The description of the details of this module is not in the scope of this paper.

4 Conclusion

There is a big body of research that suggests the importance of trust in any kind of relationship, especially between team members [16]. At the moment, the trust cognitive model by Vinanzi et al. [24] is able to perform trust evaluations based solely on past and present experiences of the involved agent, but the implementation of both the EgES and the GW seem to be able to enhance this decision making process by taking into account other factors that would influence human beings engaged in the same task, thus freeing this cognitive model from the boundaries of personal experience.

This is a position paper and the architecture described here is a design proposal, so future works will include its technological implementation and the design and execution of an HRI experiment to validate it.

To do so, a SoftBank Pepper humanoid social robot will be used. The latter is designed to operate in human environments and its interaction capabilities make it suitable for this specific scope.

The proposed work is part of a wider project which aims to build a cognitive architecture able to operate in a human-robot teaming interaction scenario where robots will act in unsupervised and dynamic contexts.

Our long-term research goal is the analysis and development of HRI systems where the artificial agent can collaborate and cooperate as a peer component in a human-like fashion. The design of a wider-scale human-robot teaming system faces several issues that must be addressed to solve the general problem. We analysed some of these factors in [8], and we discussed a theoretical cognitive model in [9].

References

1. Baars, B.J.: In the theatre of consciousness. global workspace theory, a rigorous scientific theory of consciousness. Journal of Consciousness Studies **4**(4), 292–309 (1997)
2. Baars, B.J.: Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. Progress in brain research **150**, 45–53 (2005)
3. Baars, B.J.: The global workspace theory of consciousness. The Blackwell companion to consciousness pp. 236–246 (2007)
4. Boone, R.T., Buck, R.: Emotional expressivity and trustworthiness: The role of nonverbal behavior in the evolution of cooperation. Journal of Nonverbal Behavior **27**(3), 163–182 (2003)
5. Cangelosi, A., Schlesinger, M., Smith, L.B.: Developmental robotics: From babies to robots. MIT Press (2015)
6. Castelfranchi, C., Falcone, R.: Trust Theory: A Socio-Cognitive and Computational Model. Wiley Publishing, 1st edn. (2010)
7. Castelfranchi, C., Falcone, R.: Towards a theory of delegation for agent-based systems. Robotics and Autonomous Systems **24**(3-4), 141–157 (1998)
8. Chella, A., Lanza, F., Pipitone, A., Seidita, V.: Human-robot teaming: Perspective on analysis and implementation issues. vol. 2352 (2019)
9. Chella, A., Lanza, F., Seidita, V.: A cognitive architecture for human-robot teaming interaction. In: Proceedings of the 6th International Workshop on Artificial Intelligence and Cognition. Palermo (July 2-4 2018)
10. Cho, J.H., Chan, K., Adali, S.: A survey on trust modeling. ACM Computing Surveys (CSUR) **48**(2), 28 (2015)
11. Ekman, P.: Telling lies: Clues to deceit in the marketplace, politics, and marriage (revised edition). WW Norton & Company (2009)
12. Elkins, A.C., Derrick, D.C., Burgoon, J.K., Nunamaker Jr, J.F.: Predicting users' perceived trust in embodied conversational agents using vocal dynamics. In: 2012 45th Hawaii International Conference on System Sciences. pp. 579–588. IEEE (2012)
13. Erikson, E.H.: Childhood and Society. W. W. Norton & Company (1993)
14. Falcone, R., Castelfranchi, C.: The socio-cognitive dynamics of trust: Does trust create trust? In: Trust in Cyber-societies, pp. 55–72. Springer (2001)
15. Falcone, R., Castelfranchi, C.: Socio-cognitive model of trust. In: Human Computer Interaction: Concepts, Methodologies, Tools, and Applications, pp. 2316–2323. IGI Global (2009)
16. Groom, V., Nass, C.: Can robots be teammates?: Benchmarks in human–robot teams. Interaction Studies **8**(3), 483–500 (2007)

17. Jones, G.R., George, J.M.: The experience and evolution of trust: Implications for cooperation and teamwork. Academy of management review **23**(3), 531–546 (1998)
18. Mayer, R.C., Davis, J.H., Schoorman, F.D.: An integrative model of organizational trust. Academy of management review **20**(3), 709–734 (1995)
19. Patacchiola, M., Cangelosi, A.: A developmental bayesian model of trust in artificial cognitive systems. In: 2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob). pp. 117–123. IEEE (2016)
20. Paul, E.: Emotional and conversational nonverbal signals. In: Language, knowledge, and representation, pp. 39–50. Springer Netherlands (2004)
21. Rotenberg, K.J., Petrocchi, S., Lecciso, F., Marchetti, A.: The relation between children's trust beliefs and theory of mind abilities. Infant and Child Development **24**(2), 206–214 (2015)
22. Tulving, E., et al.: Episodic and semantic memory. Organization of memory **1**, 381–403 (1972)
23. Vanderbilt, K.E., Liu, D., Heyman, G.D.: The development of distrust. Child development **82**(5), 1372–1380 (2011)
24. Vinanzi, S., Patacchiola, M., Chella, A., Cangelosi, A.: Would a robot trust you? developmental robotics model of trust and theory of mind. Philosophical Transactions of the Royal Society B **374**(1771), 20180032 (2019)