Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 145 (2018) 526–532

Procedia Computer Science

www.elsevier.com/locate/procedia

Postproceedings of the 9th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2018 (Ninth Annual Meeting of the BICA Society)

A Cognitive Model of Trust for Biological and Artificial Humanoid Robots

Rosario Sorbello[a,*], Carmelo Cali[b], Salvatore Tramonte[b], Shuichi Nishio[d], Hiroshi Ishiguro[c,d], Antonio Chella[a,b]

[a] *University of Palermo, Dipartimento dell'Innovazione Industriale e Digitale (DIID), Viale delle Scienze, 90100 Palermo, Italy*
[b] *University of Palermo, Dipartimento di Scienze Umanistiche, Viale delle Scienze, 90100 Palermo, Italy*
[c] *Intelligent Robotics Laboratory, Graduate School of Engineering Science, Osaka University, Osaka, Japan*
[d] *Hiroshi Ishiguro Laboratories, Advanced Telecommunications Research Institute International, Kyoto, Japan*

Abstract

This paper presents a model of trust for biological and artificial humanoid robots and agents as antecedent condition of interaction. We discuss the cognitive engines of social perception that accounts for the units on which agents operate and the rules they follow when they bestow trust and assess trustworthiness. We propose that this structural information is the domain of the model. The model represents it in terms of modular cognitive structures connected by a parallel architecture. Finally we give a preliminary formalization of the model in the mathematical framework of the I/O automata for future computational and human-humanoid application.

Keywords: Human-Humanoid interaction; Android Robot; Cognitive Architecture; Trust; Trustworthiness.

1. Introduction

Trust has become an issue of concern to the study of technology acceptance, users' confidence in communication, transaction and e-service systems, control of computer aided and agent systems that allow switching from automation to supervision, human-robot and human-humanoid interaction [11], [10], [9], [12], [1]. There are in fact many definitions of trust in social and cognitive sciences, which range from expecting positive outcomes from another one's behavior on the grounds of the subjective probability of gains against uncertainty and risks, to bargaining for self-interests with one another, and to establishing reciprocation and cooperation. There are also many computational

* Corresponding Author: Office.: +39-09123862635; Fax: +39-09123860537.
E-mail address: rosario.sorbello@unipa.it

10.1016/j.procs.2018.11.117

models that span trust management policy in information systems, trust model for machine learning, trust computation algorithms [4], [5], [13]. In this context we assume the definition of [8] who try to capture the cross-disciplinary common features of the concept of trust. Trust is defined as the willingness to accept one's "vulnerability" for relying on the behavior of another agent, if the risks and the uncertainty that this interdependence imply are counterbalanced by the positive expectations on her intentions and actions. Accordingly trust is a composite cognitive state that serves as condition for decision making and action, rather than a particular kind of decision or behavior. In this paper we make explicit the cognitive engines implied by this definition of trust through the constructs of the theory of Heider [2]. On this account we present a model that specifies them as modules connected by a parallel architecture. The model provides thus an interpretation of the cognitive grammar of trust, that is of the units and the rules that represent what agents acknowledge as admissible and well-formed sequences of possible behavior. Finally, we give a preliminary expression of the model in the formal theory of the I/O automata to show that it describes the structural information that enable biological and artificial agents to endow trust and assess trustworthiness.

2. The Cognitive Engines of Trust

Trust is the condition for an agent to decide and select the particular behavior of relying on another one to achieve one's goal, under which the risk and the uncertainty caused by delegating the action to carry out are traded off against the expected benefits of taking advantage of the other agent's capacity. A value is assigned to that condition, which ends up as trust or distrust, once the dependence on another one has been set against the abilities and the control of the other agent on the environment, where the action has to be carried out, which may provide benefit in comparative terms. Therefore trust requires agents to process distinct kinds of information available in the social space of possible interactions. Heider provides the constructs that make explicit what these kinds of information are, how the relevant inputs are made available in social environment and how agents have access to and process them [HEID]. The principle of his theory is that for any two subjects P and O if P wants to influence somehow O, then P has to bring about changes in the environment E, which are observable for O. Conversely for that being the case, what O perceives and aims to must be accessible to P on the grounds of how O behaves in connection to the natural and man-made furniture of the environment, which is perceived by P's standpoint. The social environment, where agents interact and hold relations with one another, is a space of perceivable variability. Any action made to fulfill intentions or to influence one another brings about changes and adds variability in the environment agents share. The variability depends on the modification of the environment but also on the manifold means or movements by which agents realize actions. Accordingly, agents have to be able to factorize the variability to abstract the relevant information. They have to discount from the variability the contingent variation of the environment, the means and the movements and to abstract the changes as outcomes of actions realized by other agents. This factorization is the same cognitive function that enables agents to abstract information in ordinary perception. For example, perceiving things means abstracting the shape as the invariant preserved across the variability of forms induced by occlusion, relative motion, distance and perspective. The shape emerges as invariant because the various forms are seen as changes connected to it in different circumstances of perception. Likewise understanding actions means abstracting the dispositions of agents as the invariant preserved across the changes and the variability of many different circumstances of interaction. The dispositions emerge as invariant because changes are connected to them as actions made to achieve distinct goals by various means. Heider calls attribution the abstraction of social invariance from the perceived variability in the space of possible interactions. As in ordinary perception, attribution amounts to connecting observable variations to underlying constants under which they appears as changes governed by some rule. This theory provides an explication of the kinds of information and cognitive engines of trust. The information agents process regards the variability in the social space of possible interactions, which concerns the environment as well as agents bodies, i.e. movements, and location. The relevant inputs are made available to each agent at the scale of perception. Agents have access to and process them by attribution. This explication is theoretically economic because this mechanism of cognitive factorization is the specialization of the perceptual abstraction of invariance under the specific constraints of social interaction. The figure 1 illustrates the cognitive engines of trust. Trust is the condition in which an agent (trustor) selects the interdependence among the possible interactions of the social space to achieve a goal by relying on another agent (trustee). This condition is characterized by risks. The trustee is an autonomous agent who may defect at any time from the interaction or cheat the trustor by abusing the interdependence for self-interest. The trustor may withdraw

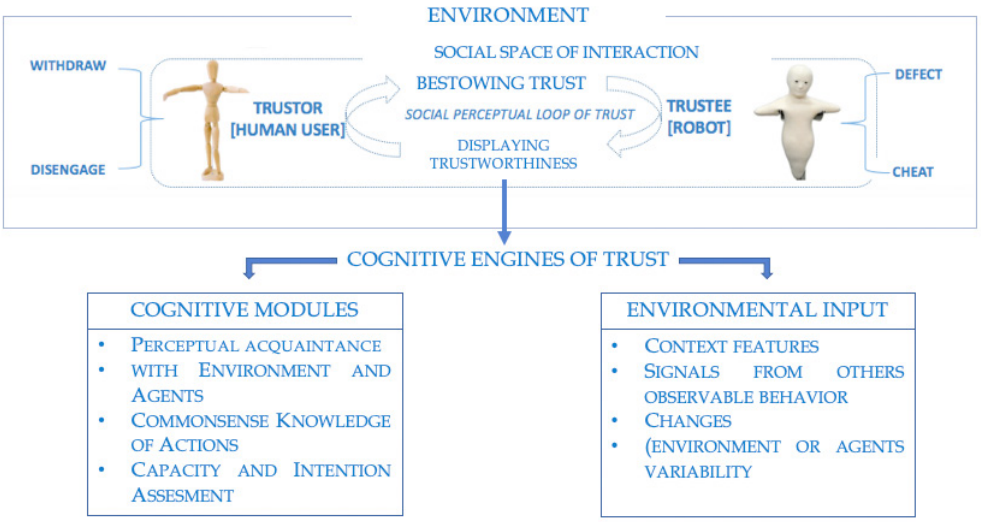

Fig. 1. The Social Perception Cycle of Trust, the model applies to any kind of agents. Here the interpretation for Human-Humanoid interaction using a Telenoid robot is suggested.

or disengage from the commitment to the interaction, because of the perceived likelihood of failure or loss due to mistakes made in assessing the capacity of the other agent. For this reason, trustworthiness is necessarily correlated to trust. From the standpoint of the trustor, however, trustworthiness is always characterized by degrees of uncertainty. The trustor cannot but have imperfect knowledge of what would make the trustee reliable and fit for carrying out actions on her behalf, because the trustor can assess this kind of information only from the observable effects of present or past actions on the shared environment. Therefore the cognitive engines of trust must provide

- the perceptual acquaintance with the environment, to which the space of interaction belongs, and the agents who share it;
- the commonsense knowledge acquired through the past and present interactions, on whose basis any agent can hold beliefs, expectations and formulate predictions on any other;
- the assessment of the capacity and intentions of the trustee on the grounds of what they have displayed through past and present interactions.

The commonsense knowledge is built on perceptual grounds. Any agent can hold expectations on any other one only if the agent has attributed a fraction of the observed variability to the dispositions of other agents rather than to contingent environmental, contextual or behavioral circumstances. Likewise any agent can predict the actions of any other only if the agent have recognized changes in the environment as actions made by other agents to get an influence over one another. Without the specialization of the abstraction of invariance, the expectations and predictions that often contribute to the assessment of trustworthiness would not be possible. The consolidation of the social commonsense knowledge works in the same way in ordinary perception. Without abstracting the invariant shape of a ball, for instance, one could not expect and predict the motion of the ball along inclined planes. For this reason bestowing trust and displaying trustworthiness form cycles in which values assigned to the condition for trust are repeatedly generated and, as a result, trust may be established or may break up as trustworthiness increases or decreases.

3. Modules and Architecture of Trust

The model of trust specifies the cognitive engines as independent modules in the sense that each module represents the abstraction of invariance under particular rules, which provides distinct domains with structures. The first module specifies the rules underlying the perception of the environment with its natural or man-made furniture. The grouping

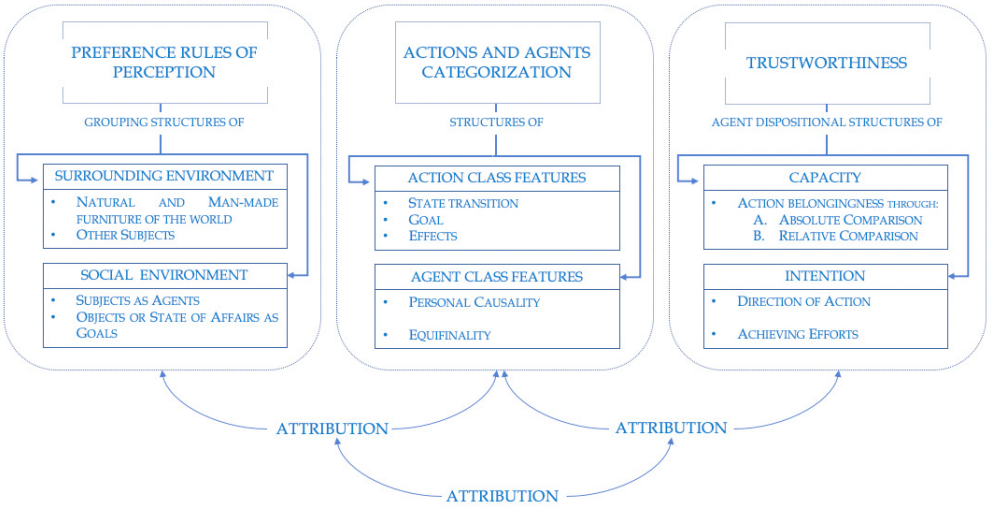

Fig. 2. The Cognitive Parallel Architecture of Trust.

mechanisms are an example of such rules, which allow subjects perceiving the world as the collection of ordered units articulated in things, events, processes, at a distance and in depth. Likewise those rules govern the perception of other subjects as moving bodies in a shared environment. The rules of perception afford agents also the units that receive the additional social meaning in the space of possible interaction where subjects are taken as agents, instead of mere autonomously moving bodies endowed with one's own viewpoint, and objects are taken as valuable in connection with goals rather than as mere things. The second module specifies the rules underlying the categorization of actions and agents, which consolidates the common knowledge on whose grounds expectations and predictions are made. The abstraction of invariant dispositions through attribution is projected onto the cognitive standards for what features changes and subjects must have to count as action and agents. In order to influence agents or to serve as observable clue of what agents perceive and aim at, any action must consist of a transition between distinct states, be directed to a goal, bring about an effect. In order to be recognized as the source of actions, any agent must show personal causality and "equifinality" [2]. Personal causality is the attribute of one agent if the changes in her surroundings across different circumstances are repeatedly observed as not contingent on any environmental source. If the variability has been so factorized that the environment sources are not the roots of it, the agent is abstracted as the cause of change, which hence is perceived as an action that falls under the range of what the agents can do. The correlate of personal causality is the belief that if one agent can do X, then he will do it. Equifinality marks out intentional from mechanical causality. Whereas for mechanical causality a change in the context where a process takes place may be necessary and sufficient to alter the outcome of the process, for personal causality a change in the context is neither necessary nor sufficient to alter the outcome of an action. As contexts vary, agents replace the means available in order to preserve action and to achieve the intended goal. The standards of actions and agents provide the structures of the social space of possible interactions. The third module specifies the rules for assessing trustworthiness. An agent assign degrees of trustworthiness to other agents as a function of the capacity and the intention, which define their reliability as trustees. The capacity and the intention enter inputs to the condition of trust as qualities of other agents in the sense that they are assessed on the basis of the what other agents display as relatively constant as tasks and circumstances vary. The capacity corresponds to a relation between the agent and the environment, whose mapping does not include what is contingent on chance and factors like good luck, tiredness that could help or hinder achieving a goal even if agents abilities were decreased or increased at will. The intention corresponds to a restriction on personal causality. An agent may bring about a change in the environment because of his physical properties or of the unintended consequences of previous actions, but a change is referred to an action if the outcome was that one the agent intended to achieve by that action. Capacity and intention build the structure of the connection between agent and actions. Each module has its rules and inputs. Unlike formal grammars, however, the rules can be qualified as "preference" rules in the sense that they don't generate all and only grammatical units, because

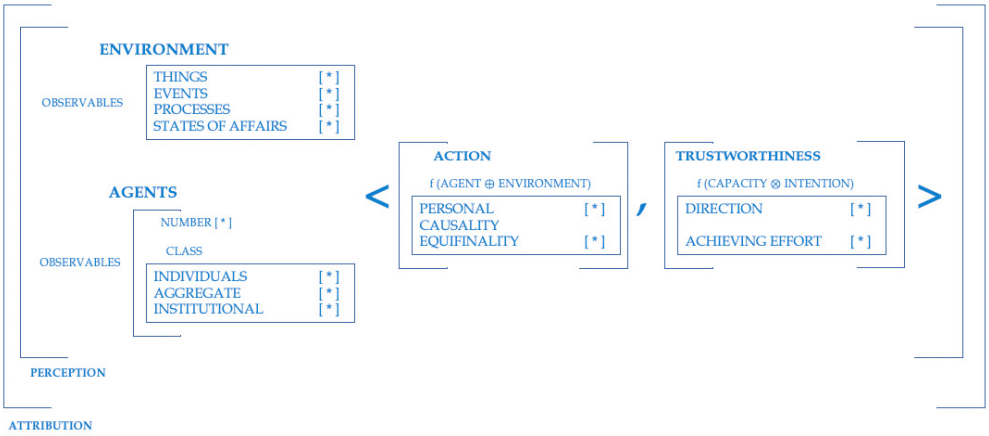

Fig. 3. Attribute-value matrix representation of the cognitive processing of trust

1. there can be more than one rule at a time to be applied,
2. rules may cooperate or compete with one another,
3. rules may be defeasible,

as it has been proved to be the case in perception. The modules are connected by attribution through a parallel architecture [3], [6], [7]. This architecture fits well the characteristics of the rules of the modules, their independence and the function of attribution. Attribution is equivalent to a production rule. However, it does not derive well-formed sequences of behavior like syntactic rules do by means of procedures of combination of primitive elements into strings of language. Rather it assembles the component units of information of trust by clipping distinct pieces of different structures. When the environment and other subjects affords variability to agents in the social space, attribution triggers the simultaneous connection of pieces of structures of the three modules. Since the pieces of structure are chunks in which rules are instanced, rather than tokens, the attribution generates well-formed units of information of trust if their connection satisfies the constraints that each chunk imposes on the other ones. The parallel architecture by which units are produced according to mutual constraints models the generative power of the cognitive grammar of trust. Figure 3 shows an attribute-value matrix description of an agent who has to solve a problem to which the condition for trust applies according to this cognitive generative grammar. It describes the implementation of attribution as one local-to-local mapping of structures on one another.

The rules and units of each module are represented in the matrix as data structures made by attributes and values. The values agent computes are assigned to the attributes, which represent the chunks of structures as instances of the rules of each modules. Such values produces well-formed information for satisfying the condition for bestowing trust or not if they meet the mutual constraints that the pieces of structures set on one another. For agents computing the information for trust means operating with the n-tuple T (Ag, Env, Obs, Act, Tw, *A) where:

- (Ag) is the class of autonomous biological or artificial agents;
- (Env) is the class of regions of the environment where social interactions may take place along with its ontology;
- (Obs) is the class of variables in (Env) perceptually accessible to elements of (Ag);
- (Act) is the class of actions carried out by elements of (Ag), which can be mapped on elements of (Obs);
- (Tw) is the class of values which are assigned to elements of (Ag) ranging over the sub-classes capacity (Can) and intention (Int);
- (*A) is the attribution.

The natural and man-made furniture of the environment along with the events or the processes, which affect or are realized in them, and the number of subjects encountered in it, be they individuals or bona fide and by fiat groups, provide the perceptual support of the space of possible interaction. In this space agents apply the rules of ordinary

and social perception, hence the corresponding classes are endowed with partial orders. In this space, whose closure is indicated by "¡" in the matrix, agents apply the specialized rule for the abstraction of social invariance. Agents abstract the elements of (Act) by (*A) as a function of (Ag + Env) in such a way that if the variability assigned to (Env) amounts to 0, in the sense that any element of (Env) does not foster or hinder an observable change, the change is referred to an action that is attributed fully to the considered agent. On the contrary if this is not the case, the change appears as something that would have occurred without any action of one agent. Values of (Env) greater than 0 may contribute to the realization of the action, but only if the value of (Ag) is not 0 in the sense that action has already been abstracted. The resulting mapping of (Act) to (Ag) requires also the determination of the personal causality and the equifinality on the basis of the records regarding (Obs). In particular, equifinality can be considered as the restrictive condition that for any x member of (Obs) and every different context $(C_1, C_2, ..., C_{(n-1)}, C_n)$ there exist a function S to a set M of means for which x is preserved. Agents abstract the elements of (Tw) by (*A) as a function of (Can x Int), in such a way that if the value of either (Can) or (Int) is 0, then the the whole value is 0. Indeed if one agent either has the capacity but not the intention or has not the capacity but has the intention, then no one would consider that agent as trustworthy since the risks of depending on her are too high. Agent define the values for (Can) in absolute or comparative terms. In absolute terms, for all elements of (Ag) if there exists one who has succeeded, or failed, in achieving the goal X associated to (Env), then the action belongs, or not, to that agent. In relative terms, for elements of (Act) if exist an agent, whose (Can) value is 0, then that action does not belong to him. The information on the trustworthiness requires also the determination of (Int), which consists of the direction (D) to a goal and the effort (E) made to achieve it. The direction informally is expressed by the conditional: if one element of (Act) is mapped by (*A) to one element of (Ag), then the latter means to achieve the goal X associated to (Env). The achieving effort is fundamental to assess the motivation of agents. On the basis of records regarding (Obs) agents assess that

- for equal values of (Can), (E) values increment as a linear function of the opposition of (Env);
- for equal opposition (Env), (E) values increment or decrement as a linear function of (Can);

Finally the matrix shows that the generative power of the cognitive grammar of trust derives from the recursive embedding of structures, represented in the matrix by the possibility that the value of an attribute is another attribute, and from the structure sharing.

4. The formalization in the I/O Automata Framework

A preliminary formalization of the model in the mathematical framework of the I/O automata is presented in figure 4 for future computational and human-humanoid application. Trust is considered as a problem that automata I/O must implement and solve in an asynchronous and distributed system of agents and communication channels. In particular Trustor and Trustee as compound automata I/Os share the same interface that is represented by the shared environment used as a channel. The three modules of the cognitive architecture provide as output the elements of a 3-pla that by means of the attribution process determines the trust level of the trustor with respect to the trustee for the specific demanded action. Cognitive modules are I/O processes and as such they involve executions and state changes.

5. Conclusions and Future Work

We presented a model of the cognitive modules of trust connected by a parallel architecture, which emerge as the specialization of the perceptual function of abstracting invariance and picking out constant properties from the observable variability of the environment under the constraints of social interaction. The model describes the cognitive engines that enable agent solving problems in the social space by means of trust as condition for selecting a particular behavior in risky and uncertain circumstances of interaction. The model represents the structural information, which is relevant to the cycle of bestowing trust and assessing trustworthiness, as the result of a cognitive grammar. The model provides an interpretation of the generative power of the cognitive functions realized by agent in commonsense experience at such an abstract level that it can be generalized over biological and artificial agents. In this connection we have proposed a preliminary formulation of the model in the framework of I/O automata. Future work will be dedicated to developing and refining the model in a formal computational language and designing a brain computer

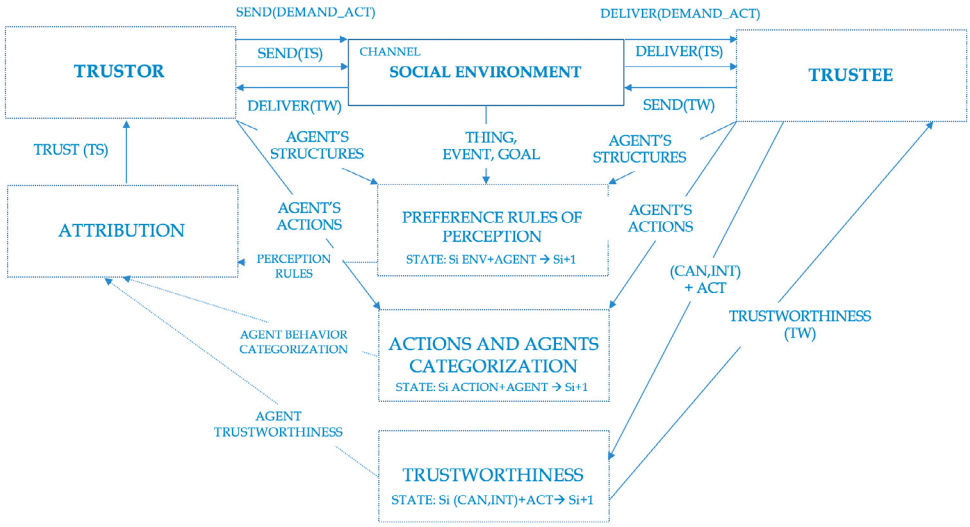

Fig. 4. I/O Automaton Framework

interface (BCI) architecture integrated with humanoid robot as experimental test generator to simulate the model for different domains of application.

References

[1] Anzalone, S., Cinquegrani, F., Sorbello, R., Chella, A., 2010. An emotional humanoid partner, in: Proceedings of the 1st International Symposium on Linguistic and Cognitive Approaches to Dialog Agents - A Symposium at the AISB 2010 Convention, pp. 1–6.

[2] Heider, F., 2013. The psychology of interpersonal relations. Psychology Press.

[3] Jackendoff, R., 2010. The parallel architecture and its place in cognitive science, in: The Oxford handbook of linguistic analysis.

[4] Krukow, K., Nielsen, M., 2007. Trust structures. International journal of information security 6, 153–181.

[5] Liu, X., Datta, A., Lim, E.P., 2014. Computational trust models and machine learning. Chapman and Hall/CRC.

[6] Pollard, C., 1996. The nature of constraint-based grammar, in: PACLIC conference, reprinted in Constructions: an HPSG Perspective, ESSLLI.

[7] Pustejovsky, J., 1991. The generative lexicon. Computational linguistics 17, 409–441.

[8] Rousseau, D.M., Sitkin, S.B., Burt, R.S., Camerer, C., 1998. Not so different after all: A cross-discipline view of trust. Academy of management review 23, 393–404.

[9] Sorbello, R., Chella, A., Calí, C., Giardina, M., Nishio, S., Ishiguro, H., 2014. Telenoid android robot as an embodied perceptual social regulation medium engaging natural human–humanoid interaction. Robotics and Autonomous Systems 62, 1329–1341.

[10] Sorbello, R., Chella, A., Giardina, M., Nishio, S., Ishiguro, H., 2016. An architecture for telenoid robot as empathic conversational android companion for elderly people, in: Intelligent Autonomous Systems 13. Springer, pp. 939–953.

[11] Sorbello, R., Tramonte, S., Giardina, M., Bella, V.L., Spataro, R., Allison, B., Guger, C., Chella, A., 2017. A human-humanoid interaction through the use of bci for locked-in als patients using neuro-biological feedback fusion. IEEE Transactions on Neural Systems and Rehabilitation Engineering PP, 1–1. doi:10.1109/TNSRE.2017.2728140.

[12] Spataro, R., Chella, A., Allison, B., Giardina, M., Sorbello, R., Tramonte, S., Guger, C., La Bella, V., 2017. Reaching and grasping a glass of water by locked-in als patients through a bci-controlled humanoid robot. Frontiers in Human Neuroscience 11. doi:10.3389/fnhum.2017.00068.

[13] Theodorakopoulos, G., Baras, J.S., 2006. On trust models and trust evaluation metrics for ad-hoc networks. IEEE Journal on selected areas in Communications 24, 318–328.