Chapter

Perspective Chapter: A Model for Measuring Trust Using BCI in Human-Humanoid Interaction

Rosario Sorbello and Carmelo Calí

Abstract

Trust is a fundamental aspect of human social interaction. With the advancement of technologies, such as brain-computer interface (BCI) systems and humanoids, arises the need for investigating human-humanoid interaction (HHI). A model to interpret BCI data in correlation to cognitive components of trust during this HHI is proposed. This will be presented by first introducing the scope of the trust in social behavior and its role as a cognitive tool for social competence. Second, a cognitive model of trust is presented with an experimental paradigm to test both general and HHI components accordingly. Then an evaluation of P300 and N400 event-related potential (ERP) signals as candidates for the neuro-cognitive markers will be performed. In particular, they will be evaluated for updating one's cognitive map and detecting a semantic violation during HHI. Finally, there will be a discussion of ERP measurement limitations as well as the prospects of incorporating electroencephalogram (EEG) oscillation (alpha, gamma, and phi) into account within the BCI system design.

Keywords: humanoid robot, human-humanoid interaction (HHI), trust, brain-computer interface (BCI), event-related potential (ERP), social perception

1. Introduction

In this paper, we argue that the research on the cognitive elements of trust brings about important insights for the study of the interaction of human subjects with artificial systems and robots. From the standpoint of the study of cognition, trust can be considered as a general-purpose cognitive ability that subserves the social competence, which is needed to solve problems of interaction among subjects in everyday life. Putting one's trust in one another as well as recognizing the trustworthiness of other people is a fundamental cognitive ability and a building block of social life [1]. In fact, the social environment is growingly populated by artificial systems and robotic devices, and trust is being increasingly discussed in the literature on artificial intelligence (AI), robotics, and automation as a fundamental factor to promote human interaction in a natural way. We make the argument that human-humanoid interaction (HHI) is a meaningful and a challenging case in point. Humanoid robots may have a profound impact on health, caregiving, work, education, and entertainment [2].

Therefore, among many interesting applications of the concept of trust in the fields of AI and robotics, we choose to focus on HHI as a test bed for the assumption that trust is required for a social interaction to take place properly with artificial agents.

The research on the elements of trust and its application can be carried out by theoretical and experimental means. In fact, in social science, the constructs of trust differ [1, 3]. Accordingly, it comes as no surprise that in the relevant literature on artificial systems, trust is assumed according to many and varied definitions, which are often biased by its intended application. Trust is mainly construed in terms of policies for authorization and delegation, and credentials management, even if attempts have been made at formalizing the human sense of trust in computational terms [4–6]. First, we present a conceptual model of the cognitive engine of trust on the assumption that the engine is made of modules recruited from independent cognitive abilities. The connection of those modules is required to solve the problems human agents face in the social space of interaction, given their limited cognitive resources. Accordingly, we interpret trust as a composite general-purpose capacity, which however can be specialized for the particular subspace of social interaction between human and robotic agents. Secondly, we describe an experimental paradigm for the application of the model to HHI, which involves the design of an interaction scenario, and the measurement of event-related potential(ERP) waves through a passive and noninvasive brain-computer interface(BCI) system, which is getting widespread in human-robot interaction(HRI) and HHI [7–12]. In the prospect of carrying out the experimental work in the near future, we present an interpretation of P300 and N400 to justify their correlation with the cognitive processing implied by the cognitive model and discuss some measurement issues that may be relevant to the experimental research.

2. The scope of trust in social behavior

Social behavior requires facing complex tasks. When deciding to take an action, agents may have incomplete knowledge of the context in which it is going to be carried out and of the other people who turn out to be involved or affected by its consequences (The term “agent” is the most general way of referring to biological or artificial entities, such as robots or humanoid robots, which are considered only from a cognitive point of view). Following Simon’s [13], incomplete knowledge can be described in terms of “bounded rationality.” If the agents were endowed with unlimited cognitive resources, overall access to transparent and relevant information, unrestricted computational capacity, and social behavior would be governed by “maximization” rules. That not being the case, maximization imposes huge costs to list alternative choices for a social task, to compare them all, to evaluate their outcomes, and to rank them accordingly as an ordering of preferences with a definite topology [14]. The term “satisficing” was introduced by Simon [13] to distinguish heuristic procedures that allow the most satisfactory solution under the given conditions, without ensuring that it is the best, from optimization procedures. In fact, agents abide by a “satisficing” criterion, rather than optimization. To make decisions and carry out actions, they follow criteria on whose basis they choose a conduct that is “good enough” under the constraint of the trade-off between the cost and the quality of that choice. Choices are made by a sort of procedural reasoning, by which the search for a good decision or action satisfies a minimal condition of acceptability, which does not demand that it is unique.

In the social domain, knowledge may be incomplete in many ways. Agents may have only partial knowledge about the available means to pursue their ends at a time or about the consequence of choices at different temporal scales. In such cases, social behavior present them with a problem that might be uniquely resolvable only if they could list all available means at a definite time or could obtain in advance information on the circumstances following that one in which they made a choice. Agents may also lack knowledge. The agent needs to make a choice about how to deal with the consequences of other people's decisions and actions. The agent is presented with a problem that might be resolvable in multiple ways, in the sense that there are many respects under which the problem could be decomposed and solved. In such a case, it is undetermined what further sources of information should be accessed by the agent to recover the knowledge needed to address the problem. Agents may even have such a knowledge that they are unable to compare properly the costs and benefits of one's own and other peoples' decision and actions if, for instance, the outcomes are not on an equal footing. Therefore, social behavior would present them with a problem, which may not be resolvable, in the sense that it cannot be decomposed into terms, which belong to common sets of features. Incomplete knowledge affects social behavior, when social tasks involve interactions within and among groups, or require adjustments at various scales and along many directions, as studies in group dynamics and social network analysis show [15, 16]. Finally, there can be extreme cases of incomplete knowledge, such as the so-called moral dilemmas. If alternative decisions and courses of action fulfill opposite common norms or explicit rules, the solution of the problem includes choices that are actually not commensurable for agents [17, 18]. Due to incomplete knowledge, social behavior carries risk and uncertainty for agents at various degrees. The variance of outcomes of social behavior is a matter of concern respectively for the predictability of their probability distribution as well as of the consideration of types of outcomes to be expected in given contexts. It is worth noting that when agents have incomplete knowledge of the types of outcomes, it is not feasible to increase the amount of information to reduce the complexity of the social task [19].

3. Trust as a cognitive tool for social competence

To solve the problems of social behavior, agents build a social competence that is supported by many cognitive abilities and pro-social skills, which range from perceiving the interdependence of events and changes of state to classifying and recognizing actions and agency roles, attributing intentions, beliefs, and desires, employing gaze contact and proxemics, and recognizing and regulating emotions. Trust plays a fundamental role in social cognition. It is the cognitive tool by which agents decompose the problems of social behavior and allocate the cognitive, computational, and information resources to other agents, by relying on or even delegating them to achieving a goal. Given the limitations of bounded rationality, putting one's trust on other people allow agents to offload the burdens of incomplete knowledge and to cope with risk and uncertainty. However, trust does not allow the ability to supplement knowledge, avoid risk, and reduce uncertainty completely. Trusting implies becoming dependent on another agent. Hence, the evaluation of the trustworthiness of other people is actually needed, to draw cues of people's reliability from overt behavior.

We submit that trust is a conditional state in which an agent weakly prefers to rely on or to delegate another agent to carry out the tasks and sub-tasks of social behavior.

We intend weak preference in a standard manner. If x, y denote elements of the set C of comparable choices for an agent a, $x \succ y$ denotes that x is preferred to y, and $x \sim y$ the relation of indifference, then:

$$x \succcurlyeq y \leftrightarrow x \succ y \vee x \sim y \text{ means that agent } a \text{ weakly prefers } x \text{ to } y. \quad (1)$$

Trusting another agent is weakly preferred as a function of:

1. The knowledge available to agents on
   a. The space of social interaction
   b. The class of action to carry out
   c. The type of outcomes,
2. The evaluation of the trustworthiness of the other agent meets the constraints put by the knowledge of the preceding domains.

This interpretation differs from the assumption that trust is a particular kind of decision or action. The domain of social behavior can be represented abstractly as the open space, whose boundary is the environment that all agents share. The environment includes everything that can be counted as a means or the objects and the states of affairs, which may be the goal or the end of agents' social behavior. When agents carry out the social behavior, they enter the social space, and their interaction defines a particular subspace. Social behavior consists of choices, decisions, actions, and interactions, which all can be considered as transformations acting on the social space. The results of such transformations are transitions between states of agents as well as objects and states of affairs. The connection between results are paths across the social space. In particular, agents are free to enter distinct states through transitions in the social space in that they walk paths leading from one state to another when they decide and act. Every state transition amounts to a change in the social space. Therefore, for each agent, the whole social space is characterized by the variability of choices, decisions, actions, and outcomes. Thus, the pair of a subspace and an instance of social behavior corresponds to a problem that agents must solve to achieve their goal. Trust is not an instance of social behavior that is a transformation, but rather a conditional state in which agents may weakly prefer to rely on other agents to carry out a transformation, rather than solving it by oneself or retreating from it. For such an alternative being weakly preferred, the benefit of offloading is weighed against the risk of becoming dependent on someone else to accomplish a social task.

4. The cognitive engine of trust

The conditional state of trust is about a relation between agents and a particular subspace where the relation is obtained. The relation represents the fact that in general agents enter one among the possible subspaces when they come to interact for the choices, decisions, and actions they make. We construe this relation as binary and leave the nature of the agents and their cardinality undetermined. Agents may have biological or artificial nature, and they may be an individual or a group that has the

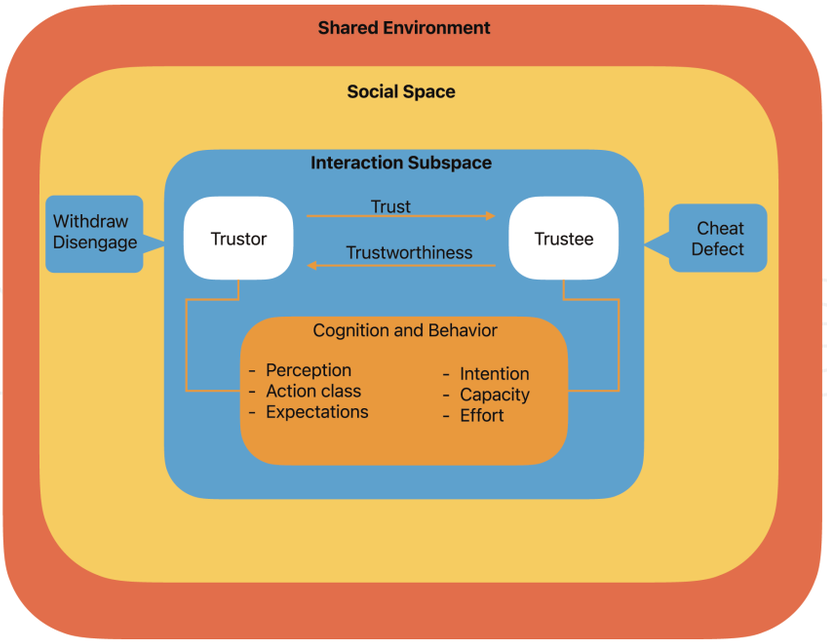

Figure 1.
The cognitive model of trust in human-human and human-humanoid interaction. Trustor and trustee are represented in terms of the relation of trust and its inverse of trustworthiness as they enter the subspace of interaction. In the boxes, underneath the required cognitive modules are listed.

title to be treated as a plural subject. In **Figure 1** a conceptual model is presented, which specifies the cognitive components that are required for the relation to occur between agents.

Figure 1 shows the trust relation and its cognitive components.

Given the set (S) of agents, the binary relation of *trust* T is such that for any a, b belonging to (S) the ordered pair (a, b) is obtained and $T(a, b)$ is true for (SxS).

$$T = \{\forall a, b \in (S) \Rightarrow (a, b) \in T \ \wedge T(a, b) = true \ \text{ for } \ (SxS)\} \tag{2}$$

The agents (a, b) paired by T are respectively the trustor a and the trustee b. The trustor is the agent who enters the conditional state of trust. If the information acquired through the cognitive modules represented underneath meet the constraint, which modules impart on the data of each of them, the conditional state is taken to be satisficing. As a consequence, the trustor *weakly prefers* to rely on or even to delegate another agent to make a decision or to carry out an action in the interaction subspace. The trustee is the autonomous agent trust is bestowed on, if the trustee displays directly or indirectly through his overt behavior to have the intention and the capacity to make the effort required by the trustor to take part in the decision and the action. The autonomy of the trustee is fundamental to trust. On the one side, relying or delegating some other agents implies that the trustor does not have control over the trustee. Otherwise, no offload of the cognitive and computational burden would take place. Autonomy is required for the trustee to supplement the trustor's incomplete knowledge. On the other side, autonomy brings in risk and uncertainty for the trustor. The trustee may cheat or defect from the interaction. Therefore, the determination of the intention, the capacity, and the effort conveys the information, which is weighed by the trustor against this kind of risk and uncertainty. For this reason, trust implies trustworthiness as its inverse relation. For any $T(a, b)$ of (SxS), *trustworthiness* is the relation:

$$T^{-1} = \{(b,a), (a,b) \in T\} \tag{3}$$

The cognitive meaning of the inverted ordered pair (b,a) is that the trustee b is endowed with trust by the trustor a.

When agents enter the subspace of interaction, the trustor needs to acquire information on the potential trustee on a direct or indirect basis, and the latter shall display the capability to act in the social space properly and suitably. Thus, the cognitive engine of trust includes distinct cognitive modules recruited for that aim. The first module is the perceptual capacity to parse the shared environment in objects, events, state of affairs, movements, and other agents, which are inherited in the social space, in which they become the elements on which social behavior is acted. In particular, this module enables an agent to observe other agents' behavior to see actions in their movements (rather than material displacements of bodies) to detect changes as outcomes of agency (rather than unintended consequences of mechanical forces). This module populates the social space, so to say. The second module is the classification of actions. It enables the agent a to attribute the changes observed in the social space to another agent b as the type of outcomes of a particular action class. The recognition of the outcome type of an action and its attribution follow a rule that can be stated intuitively thus; if the contribution of any observable element of the shared environment can be ideally reduced to zero, without canceling, hindering, or fostering a state transition, then the transition is brought about by the action of another agent that is it yields the outcome of an action. In everyday life, it may well be the case that the environment adds its effects to agents' actions, hence under this respect agency and environment hold the additive relation, in the sense that mapping a change to environmental causes may return a null value. The third module is the expectation that

1. Trustees' behavior is consistent with past and present interactions,

2. State transitions can be connected by paths in the social space.

For the time being, we may speculate that the second module is founded on the "intuitive psychology," which underlies social competence since infancy, while the third is on cognitive mechanisms akin to a partially observable Markov decision process and brain predictive maps [20–22].

The modules of the cognitive engine are correlated to the dimensions along trustworthiness of other agents assessed by the trustor. Outcomes and changes in social space are sampled by observations and assigned to agents on the basis of the perceptual and the action classification modules. The intention is extracted by the observations of actions. The capacity of an agent is inferred from type of actions instanced in particular contexts and their outcomes. If for some subspaces of interaction, the trustor observes that the outcomes of actions are invariantly obtained against or across varying environmental factors, then the agent to which they are assigned is deemed to be capable of performing that type of action. The effort is inferred from the task or the goal to be achieved. All conditions being equal, for a given capacity, the effort needs to be increased as the difficulty of the task increases. Conversely, for the difficulty to be equal, the effort needs to increase more the less capacity an agent has. For the effort to be equal, the capacity needs to be greater if the difficulty increases. Intention, capacity, and effort hold a multiplicative relation. It is sufficient that one of them falls below an expected value set by the trustor or by social standards, for the one to whom they are assigned to fail to qualify as trustworthy.

Since the modules of the cognitive engine of trust may be inherited by distinct cognitive functions, we suggest that modules are actually independent. Some abilities drawn from perception, categorization, naive psychology, and memory are recruited when agents address the solution of problems of social behavior. Once recruited, the modules act as chunks of cognitive structures, which are connected in such a way that the information acquired through different processing operations meets mutual constraints.

On the basis of such information, a partial ordering of agents in some relevant subset of S can be obtained, to which the weak preference of trusting for any agent may be anchored. If the result of feeding such information into the conditional state were below to a "satisficing" criterion, the trustor may withdraw from relying on or delegating another agent. For instance, depending on someone else turns out to be too high a cost, if not inefficient. If trust is not endowed, a social behavior does not occur, or a social interaction fails, the agents exit so to say the subspace of interaction, and go back to the social space.

5. Trust in artificial intelligence(AI) and robotics

The research on trust has important implications in AI and Robotics. In the literature, trust is listed alongside with confidence and reliability as essential factor for the use of and the interaction with artificial systems to be as efficacious and efficient as possible. As fundamental a factor of social competence as it is, trust is discussed as a construct that suggests requirements on the hardware and software design of communication and e-service systems, Internet of Things (IoT) networks, automation and cruise control systems, and interacting robots [23–29]. For instance, e-trust is defined as a dimension of e-service customers made of experienced "integrity, benevolence, and ability" is considered crucial to customer relationship management, along with the control of website interface features that are associated with user's affective states of enjoyment and anxiety [30]. This concept can be extended to e-health, peer-to-peer booking services, and to user experience of consumer electronic devices. Automation and cruise control systems design need to secure the successful collaboration of human users, by calibrating their trust, that is avoiding over-trusting and under-trusting artificial agents [31, 32]. Over-trusting leads users to overestimate agents and force their employment beyond their intended capability. Under-trusting leads users to underestimate agents. That may cause workload increase and errors for humans and disuse of the agent just in those cases in which they would perform better in automated rather than supervised mode.

Interestingly, such issues can be extended to human-robot interaction [33]. Among the artificial agents, which are increasingly populating the social space of everyday life, humanoid robots are as interesting for their impact on health, caregiving, work, education, and entertainment as challenging for the research on trust.

Humanoids are robots, which display a resemblance with humans above a theoretical lower bound that divides them from industrial and toy robots. If one assumes, as Mori et al. [34] state that the degrees of resemblance vary as the values taken by a continuous function that maps static and moving artifacts to a partially ordered subset of human aspects, a theoretical upper bound can be imaged that divides them also from stuffed pets. For their resemblance to humans, humanoids can afford opportunities for interaction and cooperation in several contexts of human activities, with a higher level of interplay than their non-humanoid robotic counterparts. Standard

questionnaires may be employed to approximate people's attitude toward humanoids such as the so-called *Goodspeed* on anthropomorphism, animacy, apparent intelligence, likability, and safety of robots and the *Nars* (Negative Attitude toward Robot Scale) on the interaction situation, social influence, and emotional engagement [35, 36]. People find humanoids acceptable and are inclined to attribute them emotional and cognitive abilities on the grounds of what they look like. A further expectation arises in human subjects that humanoids display a coherent and consistent behavior, which can be fulfilled by designing the robots in such a way that they comply with the social cues people share by tacit knowledge during their interaction in everyday contexts [37]. If this expectation was not fulfilled, a real interaction could be greatly hindered [38]. However, a particular hindrance is to be overcome with humanoids. If the degree of resemblance is plotted as the independent variable against the sense of affinity toward robots, a sudden decrease in affinity is predicted [34]. Affinity varies as an increasing non-monotone function of resemblance, in the sense that a complex sense of easiness, empathy, and familiarity grows as the resemblance initially increases from the theoretical lower bound. As a threshold of resemblance is reached, a sudden decrease of affinity is observed such that discomfort, aversion, and even eeriness are reported. Beyond the range of humanoids, the descent goes as far as to aversion toward sorts of human replicas without consciousness, such as zombies, and then affinity begins to rise again when the resemblance taken into account regards prosthetic devices.

The curve representing this decrease of affinity as a function of resemblance is known as the "uncanny valley." Even without reaching the lowest values in the range of humanoids, the uncanny valley also affects the experience with them. There are different strategies to address this problem. One is trying to reduce, if not minimize, the effect of the uncanny valley by designing humanoids that resemble humans for elementary features that are perceived as neutral as possible, that is to say, shareable by every human being, in order to generalize people's sense of acquaintance with them. Such features may range from the perception of natural parts and the kinematics of the body, which usually carry out actions or intentional expressions, to the haptic properties of the body. A case in point is the series of tele-operated Telenoid robots [39, 40]. Another one is assuming theories of interpersonal and social relationship as a framework to study and design HHI or to implement pro-social abilities in the humanoid's observable behavior [41, 42]. We suggest that the research on trust is ideally placed at the junction of those two strategies. Sorbello et al. [43] found that human subjects interacting with the Telenoid in minimally structured social scenarios do not take the robot as a screen upon which their social competence is projected or as a replica mimicking the subject who was tele-operating it. Rather, they take the Telenoid as an artificial agent that may share their environment and display pro-social abilities in its observable behavior. If the suitable subspace of interaction is provided, the Telenoid makes present what the tele-operator would do and say if he was in front of the interacting subjects. A tentative conclusion is that human subjects enclosed their interaction with the robot within a cognitive frame that inherited the social structure that ordinarily they find in human-human relations. Furthermore, they reported no negative effect of the uncanny valley because the resemblance of the humanoid was restricted to the abilities that interacting subjects found to be relevant in the intended context.

On this basis, we submit that HHI is a subspace of social space, rather than an "illusion of life" scenario [43], and that accordingly the application of the model of trust can be investigated, with the aim of specifying the conditions at which interaction can be as natural and efficient as possible.

6. An experimental paradigm to detect trust in HHI

The model of the cognitive engine of trust is testable as a whole or with respect to single modules. In this section, we describe the type of an experimental design for a pilot study to be conducted in the future. The framework of the interaction for the experiment is built in terms of the rock-paper-scissors game between two agents.

This game has been already used not only to train and test AI and robotic systems to play successfully with human partners but also to study social determinants of HRI. Ahmadi et al. [44] found that the acceptance and the attractiveness of a robotic platform are correlated to the robot's playing strategy. A significant difference was detected for a robot trained to play either randomly or according to a Markov chain model, which enables it to predict human partners' intentions. Short et al. [45] employed this game to test whether subjects recognize a robot as an agent, namely as someone who is able "to intentionally make things happen by one's action." Subjects were asked to play the rock-paper-scissors game with a robot that could either play fairly or cheat. Cheating occurred when the robot either verbally declared itself to be the winner, after several rounds in which it in fact had lost or changed its gesture after the subject's move in order to reverse the result and to turn out being the winner. By administering a questionnaire after the interaction, Short et al. [45] found that the subjects deemed the robot's behavior ambiguous when it declared to be the winner, in the sense that they were uncertain whether the robot was either affected by malfunctioning, or misread their moves, or was actually cheating. Instead, the subjects took the robot as clearly cheating, when it changed actually its move.

Although any deviation from the expected behavior was sufficient to raise their engagement in the interaction, subjects reported feeling more engaged emotionally in the latter case. When cheating was limited to verbal behavior, subjects were more inclined to seek various reasons for it, which included mechanical and processing trouble. Otherwise, when cheating was carried out by the action of changing one's move, subjects converged to attribute an intentional state to the robot, hence taking the game as a social interaction. If winning by verbal cheating was considered accidental, winning by changing the move was considered just as unfair as it could have been with other human agents.

On the basis of this result, we suggest a 2×2 factorial design for our experiment (see **Figure 2**).

The subjects will be randomly assigned to the conditions of human-human or human-robot cheating behavior. We choose to use the Telenoid because it is suitable for a Wizard of Oz experimental setup (a user-research method where a user interacts with a robot interface controlled, to some degree, by a person). Its behavior can be remotely controlled without decreasing the sense of presence and agency [46]. This allows also for controlling the manipulation of the variables and preserving the spontaneous behavior of human subjects at the same time. In this sense, the experimental setup approaches a simulation of social behavior. The experimental subjects will be asked to play the rock-paper-scissors game either with a human partner or with the robot Telenoid. In the human-human interaction condition, the human partner is blind to the aim of the experiment. One human partner will be instructed not only to play the game in the ordinary way but also to play unfairly after receiving a signal scheduled by the experimenter through a headset. In the human-humanoid interaction, the Telenoid is tele-operated to follow the same behavioral pattern. As an unfair move, we will select only changing the gesture by which the move is performed to turn out to be the winner when in reality they did not win. A verbal statement, such as

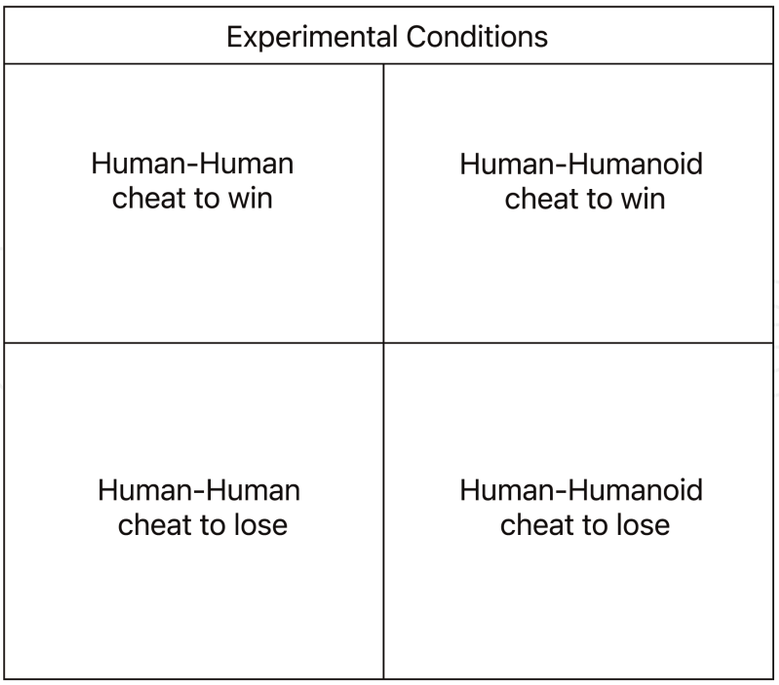

Figure 2.
This figure illustrates the 2 × 2 factorial design for the experimental study.

for instance "See, I won," may be associated, but cheating is restricted to action change. Due to the physical features of the Telenoid, the moves and their change will be visualized on a screen mounted on it. The human and humanoid cheating behavior will be scheduled to occur less frequently than the fair one. The ratio of cheat to fair moves may be respectively in the proportion of 25 or 20–75 or 80 over 100 moves. Before running the experiment, we will decide whether obtaining this ratio is a sequential or a temporal probability. However, the distribution will be such that the cheating moves will occur for the subjects in an unexpected manner, but the occurrence of cheating will always be less frequent than that of playing fairly. As factor levels, we will distinguish cheating to win and cheating to lose, instead of verbal and gesture behavior. In the first case, the robot will change its move so that it turns out to win unfairly, while in the second it will turn out to be losing the game round. If a verbal cue is present, the cheating to lose will be associated with a statement of the sort "What a pity, I've lost." This offers the advantage of focusing the subjects on a move substitution that counts as a change, which can be attributed to a choice, a decision, or an action.

We expect that if an outcome is to the detriment of oneself, it is likely to be taken as accidental rather than as a change brought about by a decision or an action. In particular, the cheating to lose behavior should be taken into account more likely as a malfunction incurring by mechanical troubles or programming errors or as a mistake made by the tele-operator, if the experiment admits to the sense of telepresence.

The HHI condition is designed to draw a baseline and to control the experimental effects, if any. The HHI condition is designed to test the application of social competence to such a specialized subspace as HHI. In general, we expect that cheating will bring about an experimental effect only if the agent is seen as otherwise trustworthy. The series of fair moves are equivalent to observations by which an agent samples a

social subspace, on whose basis the intention, the capacity and the effort of another agent are evaluated. Given the greater frequency of fair moves, which implies knowing the rules of the game, the difficulty of the task should count as constant for experimental subjects. Therefore, breaking the rule by cheating should pose the question of adjusting the trustworthiness evaluation with respect to capacity if the other agent is taken to be error prone or malfunctioning, and to intention or effort, if the other agent is taken to be playing unfairly. In both cases, a reappraisal of the benefit of trusting another agent as a partner would be needed. Only in the last case, however, this reappraisal would imply that the agent had appeared trustworthy at the outset that is a logical implication for considering him to behave unfairly. In the HHI condition, the expected main experimental effect of cheating to win should be significantly correlated to the trustworthiness of the robot, and inversely to the fact that the humanoid appears as an agent that can be endowed with trust in social interaction.

7. The neuro-cognitive markers of trust

During the interaction, ERP measures will be taken of the waves that we consider to be reasonable candidates for the neuro-cognitive markers of trust and trustworthiness: P300 and N400. In this section, we discuss the rationale for selecting those ERP waves. The BCI system design by which measures are to be taken will be presented in the section that follows.

The P300 is a candidate for technical and theoretical reasons. First, the fair-cheating condition is designed according to the standard oddball paradigm. In a classical oddball experiment, subjects are presented with two distinct stimuli that are unevenly distributed during the session. For instance, one that amounts to 80% and the other to 20% of the stimulus sequence. If EEG signals are extracted in a time window that begins 100 ms before the stimulus presentation and lasts for 800 ms, ERP waves time locked to the onset of the stimulus of interest may be processed. If the voltage variation of ERP waves is plotted at each time point, the P300 is usually recorded as the waveform time locked to the onset of the infrequent stimulus.

Many interpretations of the P300 have been proposed to determine which brain activity it represents and what neuro-cognitive meaning it may have. P300 has been construed as the marker of resource allocation, memory-related processing, and guessing an unseen event on the basis of instructions [47–49]. This latter interpretation was suggested to interpret data from an experiment on trusting information from unknown agents.

However, a distinct meaning was proposed by Refs. [50, 51], which is consistent with the characteristics of our model. P300 can be defined as the neuro-cognitive manifestation of updating a broad representation of the overall state of the environment in which agents carry out a task. Moreover, this process would bear a strategic rather than a tactical meaning. Since the P300 onset is very often detected to occur too late to have an impact on the immediate behavioral response to the stimulus, it was deemed likely that it would be a marker of the cognitive processing needed to cope with something in the future. Given our model of the cognitive engine of trust, we submit that the P300 is the neuro-cognitive marker for updating the evaluation of agents' trustworthiness. This updating is strategic because it regards the interdependence of two agents and is demanded by the change of conduct of the agent that behaves no longer fairly and breaks intentionally the rules of the game. This is a

change that is meaningful for weighing the benefits against the risks of interaction. The N400 is the other candidate for the role of the neuro-cognitive marker of trust processing. It is an ERP wave observed with a peak at 400 ms after the onset of the event to which it is time locked. It has been correlated to the violation of semantic expectancy based on lexical competence. A weaker interpretation would have the N400 being correlated to the strength of association between words. Actually, the N400 has been construed as the manifestation of either retrieving stored conceptual knowledge associated with a word or supplementing a linguistic context with the retrieved word [52, 53].

However, the N400 was also observed for violations of expectancy due to the occurrence of unrelated or inconsistent stimuli even in nonlinguistic contexts, provided that they are meaningful. For instance, line drawings or pictures inconsistent with the sequence of preceding drawings or with pictures to be completed elicit the N400 [54, 55]. Given our model of the cognitive engine of trust, we submit that the N400 is the neuro-cognitive marker of the inconsistent behavior of the agent who breaks the rule unexpectedly. Mu et al. [56] already suggested that the N400 is a reliable marker of the violations of social norms, and Salvador et al. [57] found a N400 effect when the violation occurs if subjects are related socially to one another that is have a goal that demands coordination and feel that social norms are tight and rigid. But we submit the weaker interpretation that the N400 is the marker of the unexpected outcome by which the conditions of satisfaction of a type of action are not fulfilled. By that we mean that:

The type of outcome *W* satisfies the class of action *Z* only if any instance of *W* makes *Z* into a successful action that is an action, which reaches its standard goal.

Conditions of satisfaction of this sort can be considered as preference rules, in the sense that they command how to perform an action for it to have a social meaning, even if this command is not feasible. For instance, playing a game commands to follow the rules that bind agents' behavior to be coherent and coordinated so that the interaction of playing the game can proceed. Yet that does not rule out that an agent fails to follow or break the rules. Breaking the rule, however, does not alter the meaning of the action. Rather, it prevents the action from being properly performed and violates the satisfaction of its meaning.

In our experimental paradigm, cheating is the outcome that violates the meaning of playing the game by performing the admitted moves. It is understood as an outcome that prevents the action of playing that game from being performed successfully. It is taken as a change of conduct with respect to the preference rule for that action class. Therefore, for the human-humanoid interaction condition, we submit that the null hypothesis is that

1. there is no significant effect for the P300 and N400,

2. there is no significant difference between the cheating levels.

The rejection of the null hypotheses could be interpreted to show that subjects need to adjust their strategy of interaction because the robot breaks the preference rule to play fairly. This adjustment implies that subjects need to update their evaluation of the humanoid robot's trustworthiness. As a consequence, the updating means that the humanoid robot is treated as an agent to which the concept of trust can be intuitively applied.

8. BCI system design: measurement issues

The measurement of P300 and N400 poses some methodological challenges [58]. It is well-known that ERP wave peaks are not informative by themselves. Even measuring the amplitude of the P300 may be not so reliable a measure to obtain a significant interpretation of data. The P300 amplitude is larger when subjects devote more effort to a task, and it is smaller when the subject is uncertain of whether a given stimulus is a target. Therefore, the more difficult the task, the more resources need to be allocated by the subjects, and the greater the increase of the P300 amplitude. However, it may well be the case that the subjects become uncertain to some extent as to the relevant features of the task, and this would cause a decrease in the P300 amplitude. The matter does not get better if one considers the P300 latency. A difference in the P300 amplitude is assumed to be correlated to the occurrence of a rare stimulus in a sequence of frequent stimuli or in our experimental paradigm to the cheating condition. However, it is required that the category of the stimulus for a particular task has been already recognized by the subjects, for that difference to arise. This means that any feature that increases the time to recognition would also cause a shift of the latency of the P300. Besides, the choice of measuring the P300 and the N400 as markers of trust involves another difficulty. These waves with opposite polarity succeed one another in the recording window. A greater negative deflection in the time span of the N400 might point to the higher amplitude of the N400, but equally likely to the lower amplitude of the preceding P300 wave. The reverse holds for plotting a smaller negative deflection. This means that if the subjects find playing the game with a humanoid more difficult than playing with a human partner, or if they become uncertain as to the moves of the humanoid, the P300 may be affected in such a way that it is not easier to disentangle. Furthermore, the effect that the P300 may have on the N400 both for amplitude and latency measurement might affect the correct inference of the cognitive meaning of the observed effect. To address these issues, we suggest measuring the difference waves of ERP waveforms across the conditions and employing the rectification technique to infer the interpretation from the amplitude of the recorded data [59, 60].

A difference wave is constructed by simply subtracting the voltage from different ERP waveforms at corresponding time points. After preprocessing, filtering, and removing the artifacts, we will subtract the voltage of the waveforms recorded for cheating trials from those recorded for playing fair trials across the humanoid and human conditions (see **Figure 3**).

Thus, we obtain the difference in voltage between trial types at each point in time. This procedure allows us to isolate the effect that marks the strategic updating of the trustworthiness of the humanoid. It is important to notice that the difference waves are obtained from the grand average of P300 waves computed first across trials for each subject at each electrode site and then averaged at each electrode across subjects. With a standard 10–20 system, we can try to focus on the relevant subset of the ERP components across conditions, thus increasing the likelihood to make accurate conclusions. For instance, it is known that the P300 can encompass many components. The P3a is maximally distributed in frontal electrodes and is correlated to the occurrence of an unpredictable event. The P3b is more distributed in parietal electrodes, and is correlated to a rare event that is task relevant [61]. Furthermore, this subtractive method allows selecting a wide time window (from 250 to 800 ms), and thus avoids the bias of choosing a measuring time window just in the temporal proximity of the expected ERP wave. The rectification technique is useful to isolate the N400

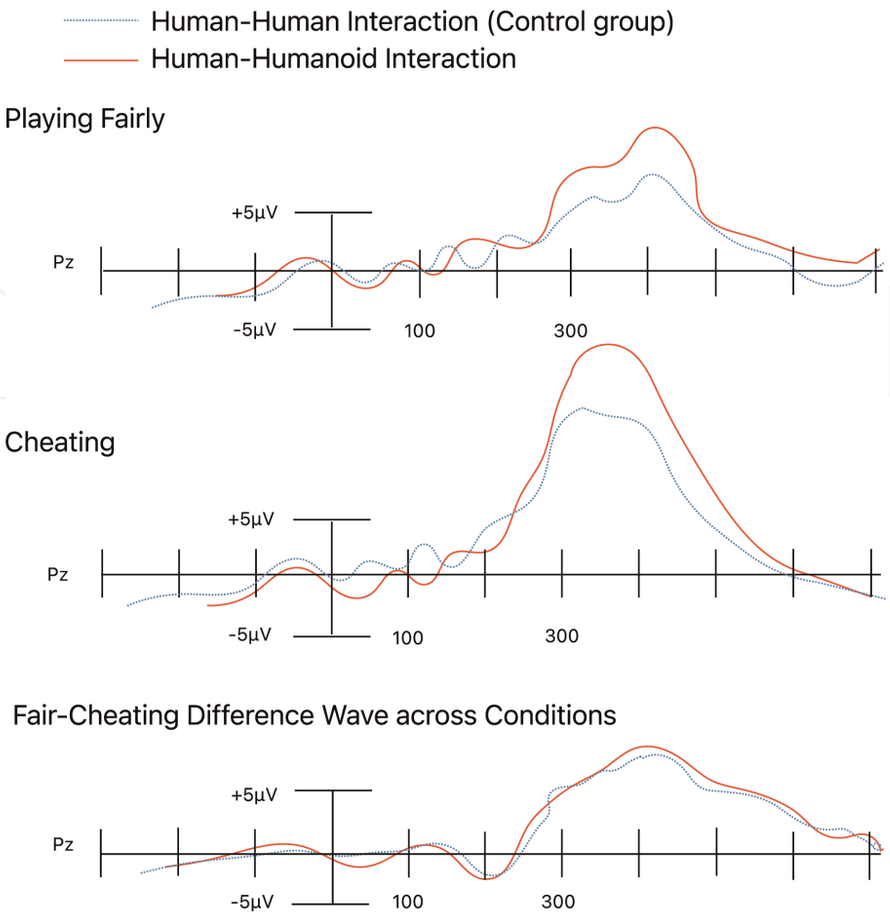

Figure 3.
Theoretical curves for the experimental effect of P300 for a fictional recording of voltages at the Pz referent point located rostrally on the scalp over the midline sagittal plane.

waveform from this extended time window of data recording, notwithstanding it is in close succession to the P300. Since P300 and N400 have opposite polarity, their effects would be canceled out were their mean amplitudes computed. Instead, one can compute the integral at the time points of the wave going from positive to negative values of the mean amplitude, just dividing the results for the duration of the considered time interval (see **Figure 4**).

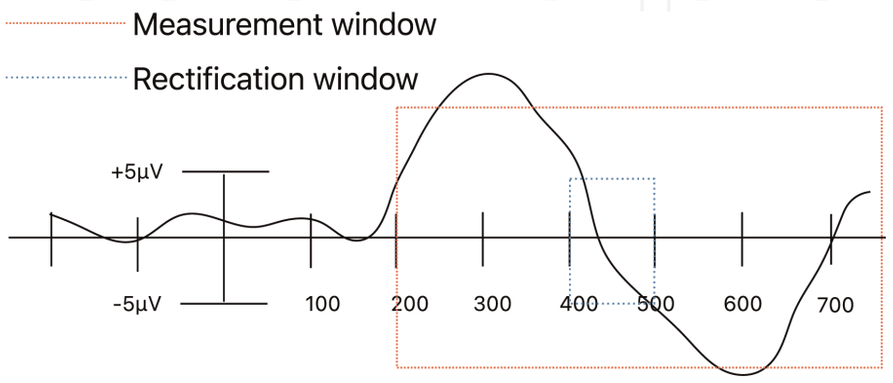

Figure 4.
Theoretical curves for the experimental effects of P300 and N400 with the smaller measurement window for rectification.

Thus, it is possible to take the amplitude value of the positive and negative areas, without cancelation and measuring what varies across conditions for the components of P300 and N400. A *t*-test will confirm if the difference across conditions is statistically significant.

To implement with the robot the experimental paradigm and to measure the ERP waves during HHI, we devised the following BCI system (see **Figure 5**).

Through the available control and command system, the tele-operator enables the Telenoid to perform the fair and the cheating moves while playing with a human subject according to the experimental design. The tele-operator is blind to the experimental hypothesis. The BCI system supplies a direct communication channel from the human subject to the raw data acquisition and storage drive.

The BCI system directly converts the brain's activity of human subjects into digital signals; in particular, the EEG signals electrical activity on the scalp as raw signals are extracted and amplified in real time from the brain of the user. During the first training phase, EEG signals are calibrated. Then, during the experimental phase, raw EEG data are acquired, amplified and stored to be further preprocessed for the abovementioned measures to be taken.

For the future research, we also plan to implement the BCI system as an online tracker (see **Figure 6**) of the neuro-cognitive markers of trust and trustworthiness in a closed human-humanoid/tele-operator loop, with an additional AI module. Among the AI alternatives for classifiers, we choose to employ a support vector machine(SVM) classifier [62–64]. We choose an SVM to classify ERP waveforms and action classes because of its relative flexibility that is provided by a nonlinear classifier and higher scalability to particular classification problems. The selection of the kernel will depend on testing its application to time series data, which would rule out kernels that yield a symmetric positive semi-definite (PSD) matrix, at least for the time being. The AI module can be trained through ERP grand averages to reduce time and errors of parameter calibration with data drawn from different individuals and across conditions [65, 66]. We submit that this choice is preferable to deep network algorithms because of the amount of training data such algorithms usually require. That would make it easier to design the experimental setup in ecologically valid conditions [67, 68]. For this online BCI system, we devised the following experimental routine. Once the EEG signals have been classified, they are fed to the SVM classifier in order to perform the training and

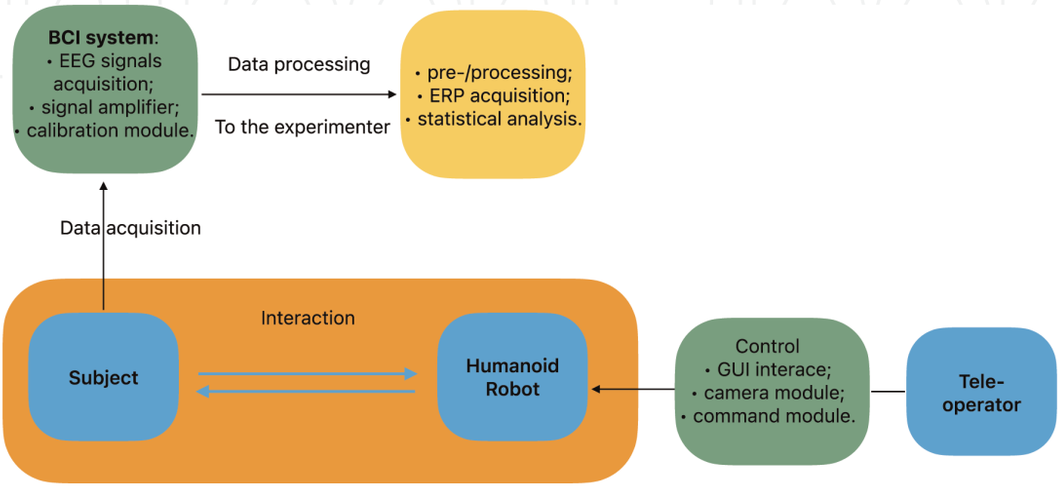

Figure 5.
The BCI system that was devised in order to measure ERP waves during HHI.

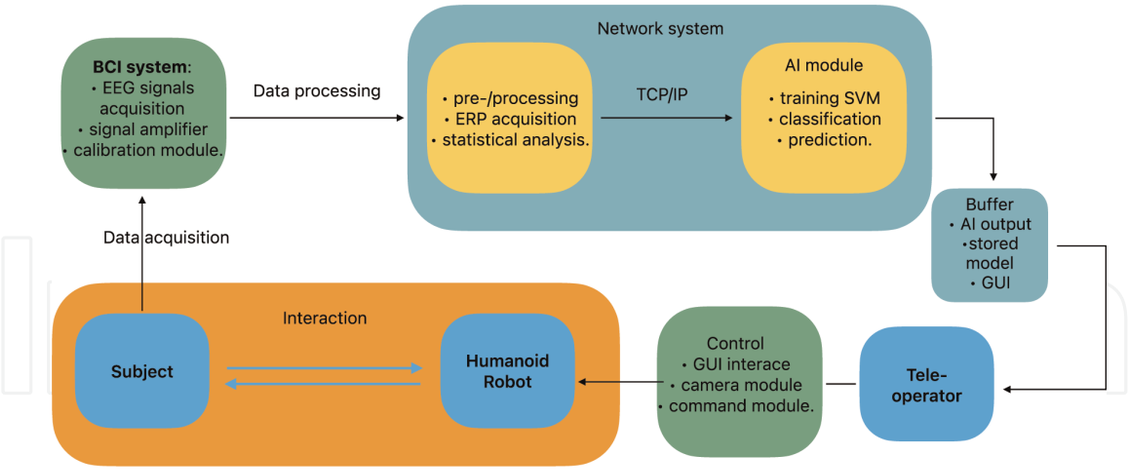

Figure 6.
The proposed BCI system as an online tracker.

the prediction phase. The data obtained from the prediction are then sent to the tele-operator, who has access to a lookup table, whose values can be used to modify the Telenoid's behavior according to the reactions of the human subject.

Finally, in the prospect of future research ERP recording data can be supplemented with event-related oscillation (ERO) measures, for which such an interpretation can be provided that makes them consistent with our model. A case in point is alpha and gamma oscillations [69]. The alpha oscillations at 10 Hz are usually correlated to recruiting multiple neuronal modules that become functionally connected, and their power in the higher frequency band range is positively correlated to cognitive tasks. Future research may be devoted to searching for their correlation to the brain activity underlying the connection of the modules the cognitive engine of trust is made of. The gamma oscillations at 40 Hz are usually related to feature binding and perceptual switching. One may hypothesize that those waves play the role of markers of the brain activity underlying the perception and action modules of trust. Phi oscillations at 9–11 Hz have been recently linked to social coordination problems [70]. These components are generated at locations above the right centro-parietal cortex and are related to either the coordination of agents' behavior or to its suppression on the basis of the observation of another agent's behavior. For the localization of the waves generator, Tognoli et al. [70] submit that those waves are linked to the activation of the mirror neuron system that is deemed to embody agents intentions and basic actions competence. Future research may be devoted to searching for their correlation to the activation of the conditional state of trust.

9. Conclusions

Trust is a building block of social life and subserves solving the problems of interaction, enabling agents to offload cognitive costs due to complexity, uncertainty, and incomplete knowledge at the price of becoming interdependent with another agent, on whom to rely on or to whom delegating decision-making and action. As such, we have claimed that trust is a general-purpose cognitive ability or a pro-social capacity that it is reasonable to be inherited in specialized contexts, such as those that involve the interaction with an artificial agent. In particular, trust can become an

ingredient that makes HRI and HHI more efficient if the robot is endowed with the level of trust that is the necessary basis for dealing with everyday experience. We have proposed a model of the cognitive engine of trust. Trust implies the recruitment of distinct cognitive modules, by which knowledge is drawn about the observable behavior of other agents, the type of actions and outcomes admitted in social spaces, and the trustworthiness of other agents. Trustworthiness is the inverse relation of trust and corresponds to the evaluation of the capacity and the effort of an agent that is someone who can bring about a change in the social space through actions, regardless of being a biological or an artificial one. From the cognitive viewpoint, trust requires a perceptual, an action classification, and an expectancy module. We submitted the paradigm and the design for a future pilot experiment to collect data on the neuro-cognitive markers of trust and trustworthiness, with particular reference to an HHI setup. We suggested the interpretations of well-known ERP waveforms (P300, N400) according to which those waves are correlated to the model, respectively as the strategic updating of a change of state of the social space of interaction and as the recognition that the conditions of satisfaction of an action class are not satisfied. We have discussed some ERP measurement problems, which may occur with P300 and N400, and presented the methodological concepts of the difference waves and the signed area technique. We suggest that the research in HHI with BCI systems can take advantage of these methods, which may also reduce the inter-subjects associativity and the calibration times. Finally, we showed the prospects of extending the BCI system of our pilot study to a complete online tracker system with a human loop for ongoing HHI as well as testing the correlation with other types of ERO oscillations and ERP waves for the future research.

Conflict of interest

The authors declare no conflict of interest.

Author details

Rosario Sorbello[1*†] and Carmelo Calí[2†]

1 Engineering Department, University of Palermo, Palermo, Italy

2 Department of Humanistic Science, University of Palermo, Palermo, Italy

*Address all correspondence to: rosario.sorbello@unipa.it

† These authors contributed equally.

References

[1] Castelfranchi C, Falcone R. Trust Theory. A Socio-Cognitive and Computational Model. Chichester, UK: John Wiley and Sons; 2010. 369 p. DOI: 10.1002/9780470519851

[2] Giardina M, Tramonte S, Gentile V, Vinanzi S, Chella A, Sorce S, et al. Conveying audience emotions through humanoid robot gestures to an orchestra during a live musical exhibition. Adv. Intell. Syst. Comput. 2018;**611**:249-261. DOI: 10.1007/978-3-319-61566-024

[3] Rousseau DM, Sitkin SB, Burt RS, Camerer C. Not so different after all: Across-discipline view of trust. Academy of Management Review. 1998;**23**: 393-404

[4] Krukow K, Nielsen M. Truststructures. International Journal of Information Security. 2007;**6**:153-181

[5] Liu X, Datta A, Lim EP. Computational Trust Models and Machine Learning. Boca Raton, FL: Chapman and Hall/CRC; 2014

[6] Song R, Korba L, Yee G. Trust in E-Services: Technologies, Practices and Challenges. Hershey, PA: IGI Global; 2007. DOI: 10.4018/978-1-59904-207-7

[7] Alimardani M, Hiraki K. Passive brain-computer interfaces for enhanced human-robot interaction. Frontiers in Robotics and AI. 2020;**7**. DOI: 10.3389/frobt.2020.00125

[8] Alimardani M, Nishio S, Ishiguro H. BCI-teleoperated androids; A study of embodiment and its effect on motor imagery learning. In: 2015 IEEE 19th International Conference on Intelligent Engineering Systems (INES). New York, NY: IEEE; 2015. pp. 347-352. DOI: 10.1109/INES.2015.7329753

[9] Beraldo G, Antonello M, Cimolato A, Menegatti E, Tonin L. Brain-computer interface meets ROS: A robotic approach to mentally drive telepresence robots. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). New York, NY: IEEE; 2018. DOI: 10.1109/ICRA.2018.8460578

[10] Ehrlich S, Wykowska A, Ramirez-Amaro K, Cheng G. When to engage in interaction—And how? EEG-based enhancement of robot's ability to sense social signals in HRI. In: 2014 14th IEEE-RAS International Conference on Humanoid Robots (Humanoids). New York, NY: IEEE; 2014. pp. 1104-1109. DOI: 10.1109/HUMANOIDS.2014.7041506

[11] Kirchner EA, Kim SK, Straube S, Seeland A, Wöhrle H, Krell MM. On the applicability of brain reading for predictive human-machine interfaces in robotics. PLoS ONE. 2013;**8**:e81732. DOI: 10.1371/journal.pone.0081732

[12] Szafir D, Mutlu B. Pay attention!: Designing adaptive agents that monitor and improve user engagement. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York, NY: ACM; 2012. DOI: 10.1145/2207676.2207679

[13] Simon HA. A behavioral model of rational choice. The Quarterly Journal of Economics. 1955;**69**(1):99-118. DOI: 10.2307/1884852

[14] Sen A. Collective Choice and Social Welfare. San Francisco: Holden Day, Inc.; 1970. 222 pp

[15] Aureli F, Schino G. Social complexity from within: How individuals experience the structure and organization of their groups. Behavioral Ecology and

Sociobiology. 2019;**73**:6. DOI: 10.1007/s00265-018-2604-5

[16] Musial K, Bródka P, De Meo P. Analysis and applications of complex social networks. Complexity. 2017;**2017**. Article No. 3014163. DOI: 10.1155/2017/3014163

[17] Gavrilets S. Coevolution of actions, personal norms and beliefs about others in social dilemmas. Evolutionary Human Sciences. 2021;**3**:E44. DOI: 10.1017/ehs.2021.40

[18] Winter F, Rauhut H, Helbing D. How norms can generate conflict. In: Jena Economic Research Papers. Jena, Germany: Friedrich-Schiller-University Jena; 2009. p. 87

[19] Botelho C, Fernandes C, Campos C, Seixas C, Pasion R, Garcez H, et al. Uncertainty deconstructed: Conceptual analysis and state-of-the-art review of the ERP correlates of risk and ambiguity in decision-making. Cognitive, Affective, & Behavioral Neuroscience. 2023;**23**:371-390. DOI: 10.3758/s13415-023-01101-8

[20] Spelke ES, Kinzler KD. Core knowledge. Developmental Science. 2007;**10**(1):89-96. DOI: 10.1111/j.1467-7687.2007.00569.x

[21] Kaelbling LP, Littman ML, Cassandra AR. Planning and acting in partially observable stochastic domains. Artificial Intelligence. 1998;**101**(1):99-134. DOI: 10.1016/s0004-3702(98)00023-x

[22] Stachenfeld K, Botvinick M, Gershman S. The hippocampus as a predictive map. Nature Neuroscience. 2017;**2017**(20):1643-1653. DOI: 10.1038/nn.4650

[23] Madsen M, Gregor SD. Measuring human-computer trust. In: Proceedings of the 11th Australasian Conference on Information Systems. Brisbane, QLD: Griffith University; 2000

[24] Chien SY, Lewis M, Semnani-Azad Z, Sycara K. An empirical model of cultural factors on trust in automation. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2014;**58**(1):859-863. DOI: 10.1177/1541931214581181

[25] Long SK, Sato T, Millner N, Loranger R, Mirabelli J, Xu V, et al. Empirically and theoretically driven scales on automation trust: A multi-level confirmatory factor analysis. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2020;**64**(1):1829-1832. DOI: 10.1177/1071181320641440

[26] Drnec K, Marathe AR, Lukos JR, Metcalfe JS. From trust in automation to decision neuroscience: Applying cognitive neuroscience methods to understand and improve interaction decisions involved in human automation interaction. Frontiers in Human Neuroscience. 2016;**10**:290. DOI: 10.3389/fnhum.2016.00290

[27] Hancock PA, Billings DR, Schaefer KE, Chen JY, De Visser EJ, Parasuraman R. A meta-analysis of factors affecting trust in human-robot interaction. Human Factors. 2011;**53**(5):517-527

[28] Esterwood C, Robert LP. The theory of mind and human-robot trust repair. Scientific Reports. 2023;**13**(1):9877. DOI: 10.1038/s41598-023-37032-0

[29] Christoforakos L, Gallucci A, Surmava-Große T, Ullrich D, Diefenbach S. Can robots earn our trust the same way humans do? A systematic exploration of competence, warmth, and anthropomorphism as determinants of

trust development in HRI. Frontiers in Robotics and AI. 2021;**8**:640444. DOI: 10.3389/frobt.2021.640444

[30] Yujong H, Dan JK. Customer self-service systems: The effects of perceived Web quality with service contents on enjoyment, anxiety, and e-trust. Decision Support Systems. 2007;**43**(3): 746-760. DOI: 10.1016/j.dss.2006.12.008

[31] Lee JD, See KA. Trust in automation: Designing for appropriate reliance. Human Factors. 2004;**46**(1):50-80. DOI: 10.1518/hfes.46.1.5030392

[32] Lebiere C, Blaha Leslie M, Fallon CK, Brett J. Adaptive cognitive mechanisms to maintain calibrated trust and reliance in automation. Frontiers in Robotics and AI. 2021;**8**:652776. DOI: 10.3389/frobt.2021.652776

[33] Ullrich D, Butz A, Diefenbach S. The development of overtrust: An empirical simulation and psychological analysis in the context of human–robot interaction. Frontiers in Robotics and AI. 2021;**8**. Article No. 554578. DOI: 10.3389/frobt.2021.554578

[34] Mori M, MacDorman KF, Kageki N. The uncanny valley [from the field]. IEEE Robotics and Automation Magazine. 2012;**19**(2):98-100. DOI: 10.1109/MRA.2012.2192811

[35] Bartneck C, Kulić D, Croft E, Zoghbi S. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. International Journal of Social Robotics. 2009;**1**(1): 71-81. DOI: 10.1007/s12369-008-0001-3

[36] Nomura T, Suzuki T, Kanda T, Kato K. Measurement of negative attitudes toward robots. Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems. 2006;**7**(3):437-454. DOI: 10.1075/is.7.3.14nom

[37] Vianello L, Penco L, Gomes W. Human-humanoid interaction and cooperation: A review. Current Robotics Reports. 2021;**2**:441-454. DOI: 10.1007/s43154-021-00068-z

[38] Anzalone SM, Boucenna S, Ivaldi S, Chetouani M. Evaluating the engagement with social robots. International Journal of Social Robotics. 2015;**7**(4):465-478

[39] Ishiguro H, Nishio S, Chella A, Sorbello R, Balistreri G, Giardina M, et al. Perceptual social dimensions of human - humanoid robot interaction. In: Lee S, Cho H, Yoon KJ, Lee J, editors. Intelligent Autonomous Systems, Advances in Intelligent Systems and Computing. Vol. 12. Berlin, Heidelberg: Springer; 2013. p. 194. DOI: 10.1007/978-3-642-33932-538

[40] Sorbello R, Chella A, Giardina M, Nishio S, Ishiguro H. An architecture for telenoid robot as empathic conversational android companion for elderly people. In: Menegatti E, Michael N, Berns K, Yamaguchi H, editors. Intelligent Autonomous Systems, Advances in Intelligent Systems and Computing. Vol. 13, 302. Cham: Springer; 2016. DOI: 10.1007/978-3-319-08338-468

[41] Fox J, Gambino A. Relationship development with humanoid social robots: Applying interpersonal theories to human–robot interaction. Cyberpsychology, Behavior, and Social Networking. 2021;**24**(5):294-299

[42] Perugia G, Paetzel-Prüsmann M, Alanenpää M, Castellano G. I can see it in your eyes: Gaze as an implicit cue of uncanniness and task performance in

repeated interactions with robots. Frontiers in Robotics and AI. 2021;**8**: 645956. DOI: 10.3389/frobt.2021.645956

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

[44] Ahmadi E, Pour AG, Siamy A, Taheri A, Meghdari AF. Playing rock-paper-scissors with RASA: A case study on intention prediction in human-robot interactive games. International Conference on Software Reuse. 2019; **11512**:55-70. DOI: 10.1007/978-3-030-20898-3_5

[45] Short E, Hart J, Vu M, Scassellati B. No fair!! An interaction with a cheating robot. In: 2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Osaka, Japan. New York, NY: IEEE; 2010. pp. 219-226. DOI: 10.1109/HRI.2010.5453193

[46] Steinfeld A, Jenkins OC, Scassellati B. The Oz of Wizard: Simulating the human for interaction research. In: Human Robot Interaction, March 11–13, 2009, La Jolla, California, USA. 2009

[47] Isreal JB, Chesney GL, Wickens CD, Donchin E. P300 and tracking difficulty: Evidence for multiple resources in dual-task performance. Psychophysiology. 1980;**17**:259-273

[48] Wilding EL, Ranganath C. Electrophysiological correlates of episodic memory processes. In: Luck SJ, Kappenman ES, editors. The Oxford Handbook of Event-Related Potential Components. New York: Oxford University Press; 2012. pp. 373-395

[49] Boudreau C, McCubbins MD, Coulson S. Knowing when to trust others: An ERP study of decision making after receiving information from unknown people. SCAN. 2009;**4**:23-34

[50] Donchin E. Surprise! ... Surprise? Psychophysiology. 1981;**18**:493-513

[51] Donchin E, Coles M. Is the P300 component a manifestation of context updating? Behavioral and Brain Sciences. 1988;**11**(3):357-374. DOI: 10.1017/S0140525X00058027

[52] Kutas M, Van Petter CK, Kluender R. Psycholinguistics electrified II (1994–2005). In: Traxler MJ, Gernsbacher MA, editors. Handbook of Psycholinguistics. 2nd ed. New York: Elsevier; 2006. pp. 83-143

[53] Hagoort P. The memory, unification, and control (MUC) model of language. In: Meyer AS, Wheeoldon L, Krott A, editors. Automaticity and Control in Language Processing. Vol. 2007. Hove: Psychology Press; 2007. pp. 243-270

[54] Ganis G, Kutas M, Sereno MI. The search for "common sense": An electrophysiological study of the comprehension of words and pictures in reading. Journal of Cognitive Neuroscience. 1996;**8**(2):89-106. DOI: 10.1162/jocn.1996.8.2.89

[55] Willems RM, Özyürek A, Hagoort P. Seeing and hearing meaning: ERP and fMRI evidence of word versus picture integration into a sentence context. Journal of Cognitive Neuroscience. 2008;**20**(7):1235-1249. DOI: 10.1162/jocn.2008.20085

[56] Mu Y, Kitayama S, Han S, Gelfand MJ. How culture gets embrained: Cultural differences in event-related potentials of social norm violations. National Academy of Sciences

of the United States of America. 2015; **112**(50):15348-153453. DOI: 10.1073/pnas.15098391

[57] Salvador CE, Mu Y, Gelfand MJ, Kitayama S. When norm violations are spontaneously detected: An electrocortical investigation. Social Cognitive and Affective Neuroscience. 2020;**15**(3):319-327. DOI: 10.1093/scan/nsaa035

[58] Luck SJ. An Introduction to the Event-Related Potential Technique. 2nd ed. Cambridge, MA: MIT Press; 2014. 374 p. ISBN: 9780262525855

[59] Hansen JC, Hillyard SA. Endogenous brain potentials associated with selective auditory attention. Electroencephalography and Clinical Neurophysiology. 1980;**49**:277-290

[60] Sawaki R, Geng JJ, Luck SJ. A common neural mechanism for preventing and terminating the allocation of attention. Journal of Neuroscience. 2012;**32**(31): 10725-10736

[61] Polich J. Neuropsychology of P300. In: Luck SJ, Kappenman ES, editors. Oxford Handbook of Event-Related Potential Components. New York: Oxford University Press; 2012. pp. 159-188

[62] Aricò P, Borghini G, Di Flumeri G, Sciaraffa N, Colosimo A, Babiloni F. Passive BCI in operational environments: Insights, recent advances, and future trends. IEEE Transactions on Biomedical Engineering. 2017;**64**(7):1431-1436. DOI: 10.1109/TBME.2017.2694856

[63] Chamola V, Vineet A, Nayyar A, Hossain E. Brain-computer interface-based humanoid control: A review. Sensors (Basel). 2020;**20**(13):3620. DOI: 10.3390/s20133620

[64] Nagel S, Spüler M. World's fastest brain-computer interface: Combining EEG2Code with deep learning. PLoS ONE. 2019;**14**(9): e0221909. DOI: 10.1371/journal.pone.0221909

[65] Sprague SA, McBee MT, Sellers EW. The effects of working memory on brain-computer interface performance. Clinical Neurophysiology. 2016;**127**: 1331-1341

[66] Saha S, Baumert M. Intra-and inter-subject variability in EEG-based sensorimotor brain computer interface: A review. Frontiers in Computational Neuroscience. 2019;**13**:87

[67] Tanveer MA, Khan MJ, Qureshi MJ, Naseer N, Hong KS. Enhanced drowsiness detection using deep learning: An fNIRS study. IEEE Access. 2019;**7**:137920-137929

[68] Lotte F, Jeunet C, Chavarriaga R, Bougrain L, Thompson DE, Scherer R. Turning negative into positives! Exploiting 'negative' results in Brain–Machine Interface (BMI) research. Brain-Computer Interfaces. 2020;**6**: 178-189. DOI: 10.1080/2326263X.2019.1697143

[69] Başar E, Başar-Eroglu C, Karakaş S, Schürmann M. Gamma, alpha, delta, and theta oscillations govern cognitive processes. International Journal of Psychophysiology. 2001;**39**(2–3): 241-248. DOI: 10.1016/s0167-8760(00)00145-8

[70] Tognoli E, Lagarde J, DeGuzman GC, Kelso JA. The phi complex as a neuromarker of human social coordination. Proceedings of the National Academy of Sciences of the United States of America. 2007;**104**(19): 8190-8195. DOI: 10.1073/pnas.0611453104