Robot's Self-Trust as Precondition for being a Good Collaborator

Filippo Cantucci
ISTC-CNR
Rome, Italy
filippo.cantucci@istc.cnr.it

Rino Falcone
ISTC-CNR
Rome, Italy
rino.falcone@istc.cnr.it

Cristiano Castelfranchi
ISTC-CNR
Rome, Italy
cristiano.castelfranchi@istc.cnr.it

Institute of Cognitive Science and Technology, National Research Council of Italy, (ISTC-CNR), Rome

Abstract

In Human Robot cooperation scenarios, building a robot that can be defined a good collaborator, means endowing it with the capability to evaluate not only the physical environment, but especially the mental states and the features of its human interlocutor, in order to adapt its behavior every time she/he requires the robot's help. The quality of this kind of evaluation, underlies the robot's capability to operate a meta-evaluation of its own predictive skills to build a model of the interlocutor and of her/his goals. The robot's capability to *self-trust* his skills to interpret the interlocutor and the context, is a fundamental requirement for producing smart and effective decisions towards humans. In this work we propose a simulated experiment, designed with the goal to test a cognitive architecture for trustworthy human robot collaboration. The experiment has been designed in order to demonstrate how the robot's capability to learn its own level of self-trust on its predictive abilities in perceiving the user and building a model of her/him, allows it to establish a trustworthy collaboration and to maintain an high level of user's satisfaction, with respect to the robot's performance, also when these abilities progressively degrade.

1   The needs of collaborative robots in Human-Robot Cooperation

The complexity of the AI systems (i.e. social robots, autonomous cars, virtual assistants etc.) surrounding us is every day more demanding and requires a consequent capability of these systems to be trusted by humans, just as humans are able to do when they collaborate with each other [KS20]. In the context of *Human-Robot Cooperation*, the sense of human vulnerability due to the presence of the robot [SSTJS18], can be reduced by changing the role of the robot itself: from a passive executor, to a smart and *active collaborator* [FC01].

Let's consider the following collaborative scenario: a human X (the *trustor*) and a robot Y (the *trustee*) collaborate so that X has to trust Y, in a specific context, for executing a task τ and realizing the results that include or correspond to the X's $Goal_X(g) = g_X$ [CF10]. In this context, X relies on Y for realizing some part of the task she/he has in mind (*task delegation*); on its side, Y decides to help X, to replace her/him and perform

a sequence of actions that are included in the X's plan, in order to achieve some of her/his goals or sub-goals (*task adoption*). The capability to implement a smart task adoption distinguishes a collaborator from a simple tool, and presupposes intelligence and *autonomy* [CF98]. Being truly cooperative implies more than the simple concept of *execution* of a prescribed action. For example, in order to adopt some goal of X in an intelligent form, Y has to understand the X's mental states (i.e. goals, beliefs, expectations about Y's behavior) and it has to adjust the delegated action to the represented mental states, to the context and to its own current abilities and characteristics. In their much complex sense, cooperation and help require more autonomy and *initiative*. A real collaborative trustee should provide to the trustor different kind of help, according with [CF98]:

- *Sub help*: Y satisfies a sub-part of the delegated world-state (so satisfying just a sub-goal of X),

- *Literal help*: Y adopts exactly what has been delegated by X,

- *Over help*: Y goes beyond what has been delegated by X without changing X's plan (but including it within a hierarchically superior plan),

- *Critical-Over help*: Y realizes an over help and in addition modifies also the original plan/action (included in the new meta-plan),

- *Critical help*: Y satisfies the relevant results of the requested plan/action (the goal), but modifies that plan/action,

- *Critical-Sub help*: Y realizes a sub help and in addition modifies the (sub) plan/action,

- *Hyper-critical help*: Y adopts goals or interests of X that X itself did not take into account (at least, in that specific interaction with Y): by doing so, Y neither performs the specific delegated action/plan nor satisfies the results that were delegated. In practice, Y satisfies other goals/interests of X by realizing a new plan/action.

Y has to exploit its autonomy, competence and cognitive skills to find the better or a possible solution for X's goal. This not necessarily should require a negotiation, discussion, agreement; it might be an initiative of Y by expecting that X will understand why. This is precisely what intelligent robots must have and these are the kind of partners the humans need.

How would this advanced form of cooperation would be possible? What are some of the capabilities that a robot has to show for enhancing trust in its human interlocutor? A smart and trust-based collaboration between humans and intelligent robots requires, among many others things, complex cognitive capabilities these artificial systems must be endowed with: mental attribution, adjustable autonomy, user profiling and user behavior adaptation, behavior transparency. Besides the capabilities to evaluate the interlocutor and/or the contextual physical environment, a robot (as a trustee) should be able also to operate a *meta-evaluation*: how much itself would be able to interpret and produce the evaluations regarding the trustor? How much is reliable its capability to perceive or infer the trustor's features? On the basis of its own capabilities to perceive or to act in the world, the hypothesis or prediction it has made, the chosen course of action, are the best or the most effective, with respect to the needs, the features and the mental states of the interlocutor? Smart help has to be based on different capabilities to interpret the environment and the interacting user, but first of all, it has to be based on the robot's capability to realistically self-assess the level of trustworthiness on its ability to interpret the collaborative and potentially uncertain context, including the interacting user [HMDAR16, IAF+19]. The outcome of the meta-evaluation expressed above represents the *robot's self-trust* for adopting a delegated task. In practice, the robot uses this evaluation of its own specific abilities as a filter for their use with respect to the interlocutors with whom it is interacting. The robot learns the trustworthiness of its skills and, on the basis of the context and the task to carry out, establishes which skills to use and how trustworthy (from its point of view) will be the solution it will propose to its interlocutor. So robot's self-trust can be viewed as a precondition for exploiting the robot's interpretative skills accordingly to its own interlocutor, in order to foster a true and deep relationship of collaboration and trust with her/him.

A form of intelligent help that can provide results beyond those explicitly requested by the interlocutor implies risks. One of the possible consequences of this form of help can be the emergence of *collaborative conflicts* between the human (the trustor) and the robot (the trustee) that adopts the task, due to the robot's willingness to collaborate and to help the user better and more deeply than required. Sometimes, the difference between the results of the adopted task provided by the robot and the user's expectations, could lead the interlocutor

to a complete lack of trust towards the robot. We are not just considering the robot's failure in the precise delegated task: failures become more evident every time the robot goes beyond the delegated task and the results are too much distant (or even in conflict) from the user's expectations. Among humans these conflicts can be mitigated by the experience: humans learn to measure their competence in achieving specific results, or making the right prediction about the correctness of a chosen behaviour, on the basis of the context and the interlocutor; furthermore, on this basis, they learn to self-trust their own abilities/skills (with respect to both the interlocutors and the tasks). Similarly, robots can learn to trust their capabilities to evaluate the interlocutors (and consequently to build and use the cognitive models they attribute to them) trough a repetitive interactions with humans. For example, a robot can exploit the feedback provided by its interlocutor any time she/he delegates to it a task and receive an evaluation (i.e. user's satisfaction) on the results of the robot's adoption process.

In this work we propose a preliminary, simulated experiment, designed with the goal to test a cognitive architecture [CF20] for trustworthy human robot collaboration. The designed architecture allows a BDI robot [RG$^+$95], with its own mental states (beliefs, goals, plans and so on) to expose a wide range of cognitive skills that support an effective, smart and trustworthy collaboration, every time a human user delegates to it a task to achieve in her/his place. In particular, we focused on endowing the robot with the capabilities to i) adapt its level of collaborative autonomy, providing an intelligent help (based on the levels of help formalized in [CF98]) every time the user delegates to it a task to accomplish; the autonomy adaptation leverages on the agent's capabilities to profile the user and to have a theory of mind of her/him [DA16] ii) learn its limits in interpreting the needs of the interlocutor, by measuring its degree of self-trust on its predictive abilities in perceiving the user; the agent chooses those abilities that maximize the user's task performance evaluation. In particular the simulation aims at demonstrating how the robot's capability to learn the level of self-trust on its predictive abilities in perceiving the user, allows it to choose the best user's model (as a collection of mental states) and to preserve an high level of the user's task performance evaluation.

2 The proposed experiment

The experiment designed for testing the cognitive architecture proposed in [CF20], has been implemented by exploiting the well known multi-agent oriented programming (MAOP) framework *JaCaMo* [BBH$^+$13], that integrates three different multi-agent programming levels: agent-oriented (AOP), environment-oriented (EOP) and organization-oriented programming (OOP). Basically, the experiment simulates the process of task delegation and task adoption between a robot and multiple users, grouped in classes of users, in a specific application domain.

2.1 The experimental settings

We figured the following interactive scenario: the robot is a touristic assistant that helps people to organize different touristic activities offered by a city (i.e. eat in a restaurant, visit a museum, visit a monument, drink something in a bar, enjoy the city doing multiple daily activities). The experiment is based on the interaction between two agents: the user and the robot. Both of them are implemented as Jason [BH05] agents. The user has her/his own mental states represented in form of beliefs, goals and plans and interacts with the robot by delegating to it a task. On its side, the robot is able to represent and attribute mental states to the user and to itself and, on the basis of its capabilities to profile the user and build a model of her/him, to adopt the delegated task at different levels of help (section 1).

The experiment has been designed with the goal to show the importance for a robot to self estimate the level of trustworthiness associated to its expertise in building a profile of the interacting user. This capability lets the robot choose the best and suitable task to adopt with respect to the user's features, also when its skills progressively degrade and can be considered not trustworthy. Indeed, the robot is able to sort these skills on the basis of the corresponding level of trustworthiness, and leverage on the most trustworthy among them for deciding how to adopt the task delegated.

As mentioned above, two agents populate the simulation: the agent robot $\mathcal{R}$ and the agent user $\mathcal{U}$. The agent $\mathcal{U}$ is characterized by a profile $\mathcal{P}_{\mathcal{U}} = \{Age, Economic\ status, Category, Education\ level, Company\}$, a collection of five physical and social features. Every feature is associated to *sub-components* and real values $r_{H_i} \in [0, 1]$ belonging to specific intervals that are bonded to the sub-components. Table 1 shows the relations between features, sub-components and intervals. We decide to consider these groups of user's demographic features, because they are all concrete characteristics that help the robot, operating in a touristic domain, to narrow down which segment of population the interacting users best fit into. That means the robot can split a larger group into

Feature	Sub-component [interval]
Age	young [0, 0.33] adult [0.34, 0.66] old [0.67, 1]
Category	loco tourist [0, 0.33] foreign tourist [0.34, 0.66] resident [0.67, 1]
Economic status	low economic status [0, 0.33] medium economic status [0.34, 0.66] high economic status [0.67, 1]
Education level	low education [0, 0.33] medium education [0.34, 0.66] high education [0.67, 1]
Company	single [0, 0.33] in couple [0.34, 0.66] in family [0.67, 1]

Table 1: Map of the relations between features, sub-components and intervals

subgroups based on, for example, their educational level, age, income. This kind of physical, social and relational features are largely used, easy to collect and they are reasonably good predictors of user preferences [BER15]. For example, demographic recommendation system generate recommendations based on the user demographic attributes [MKI19, Paz99]. In our case the robot is able to filter and categorize the interacting users based on their attributes and recommends the most suitable service (restaurant, museum, monument or bar) by utilizing the chosen demographic data collected in its profile. The partition of the features into sub components is an approximation that allows the robot to cluster users into a series of discrete categories, commonly used by human for identify expected behaviors or character traits, related to that particular category [SADL18].

Users are organized into *classes of populations*: each class collects together users with the same profile (in terms of sub-components). Each user of a class distinguishes from the others due to five real values r_{H_i} for $i = 1, .., 5$ randomly picked up from the interval associated to the sub-components. The decision making system of $\mathcal{R}$ is designed following the principles described in [CF20]. The robot is able to recognize and classify, as set of specific sub-components, the features collected in $\mathcal{P}_{\mathcal{U}}$, consistent with the table 1. $\mathcal{R}$ is not always able to infer all the features of $\mathcal{U}$; that depends on the robot's *accuracy* to estimate a feature of $\mathcal{P}_{\mathcal{U}}$. In this experiment we decide to define two levels of accuracy: a low level of accuracy, that means the robot has great difficulties in distinguishing a feature, and an high level of accuracy, corresponding to the fact that it is perfectly able to recognize a feature. We have designed the simulation so that $\mathcal{R}$ can estimate the sub-components collected in $\mathcal{P}_{\mathcal{U}}$, but it is not able to perfectly recognize the real values r_{H_i} for each user; because of that, it associates to every feature it has estimated, the mean value of the corresponding intervals defined in the table 1. We observe that, if the robot profiles a feature correctly, the corresponding mean value will be close to the value r_{H_i} of the user (for that feature), while if the robot is not able to infer correctly the feature, this value will be distant from that of the user.

It is important to specify that the robot's beliefs are organized according to the features that are classified within $\mathcal{P}_{\mathcal{U}}$ and which are perceivable by the robot itself. $\mathcal{R}$ has available (among the set of its mental states) a subset of beliefs where are represented information about a finite number of services that a city offers: restaurants, museums, monuments to visit and places for having fun (night clubs, bar and so on). Each service is described with respect to the features described in table 1: for example, in the robot's beliefs base exist restaurants much more suitable to young people, instead of monuments or museums much more adapt to people with an high level of education, and so on. The robot is able to select the most suitable service with respect to the features that it has been able to infer from $\mathcal{U}$. This criterion of choice can lead the robot to select the most adapt service with respect to the user's profile or not, on the basis of its own profiling skills accuracy.

2.2 The Experiment description

The experiment is a simulation of several *trials* – interactions between $\mathcal{R}$ and 100 users belonging to the same class (population of users) – involving the robot and different users. Every interaction reproduces the mechanism of delegation and adoption: $\mathcal{U}$ delegates a task to $\mathcal{R}$ and the robot adopts the task at different levels of intelligent help, among those introduced in section 1. We defined a class of population $\mathcal{C}_1$ formed by users that have the following profile (collection of sub-components): $\mathcal{P}_{\mathcal{U}} = \{young, medium\ Economic\ Status, foreign\ Tourist, medium\ Education, single\}$. Each interaction requires that the

current user delegates to the robot the goal to *eat in a restaurant*. The request might be further specified by giving the name of the restaurant, the type of restaurant and the area of the city in which it is located. We decide to specify only the area of the city where the user desires to eat.

2.3 Building robot's self-trust

The robot $\mathcal{R}$ builds its self-trust for adopting the delegated task τ by means of a training phase, with the goal to learn the levels of trustworthiness associated to its own user profiling capabilities. The training phase requires that the robot performs an interaction with a population of a specific class formed by 100 users. Every user $\mathcal{U}$ delegates to $\mathcal{R}$ the same task (i.e. eat in a restaurant); for its part, the robot adopts the task at a literal level of help. At every interaction $\mathcal{R}$ computes a *robot's skill trustworthiness* value, each for every feature that forms $\mathcal{P}_{\mathcal{U}}$. These values depend on the feedback provided by the users during the training phase. We designed a robot that explicitly asks for feedback, once it accomplishes a task to be achieved on behalf of $\mathcal{U}$. Every question the robot asks humans aims at evaluating how the delegating user has been satisfied by the robot's task adoption; different user's satisfaction dimensions are investigated, each of them corresponding with the different abilities of the robot to profile the user. In this way $\mathcal{R}$ can evaluate how each of its skills performs (and to measure its trustworthiness) with respect to build $\mathcal{P}_{\mathcal{U}}$. Furthermore, $\mathcal{R}$ can sort the skills on the basis of the measured level of trustworthiness.
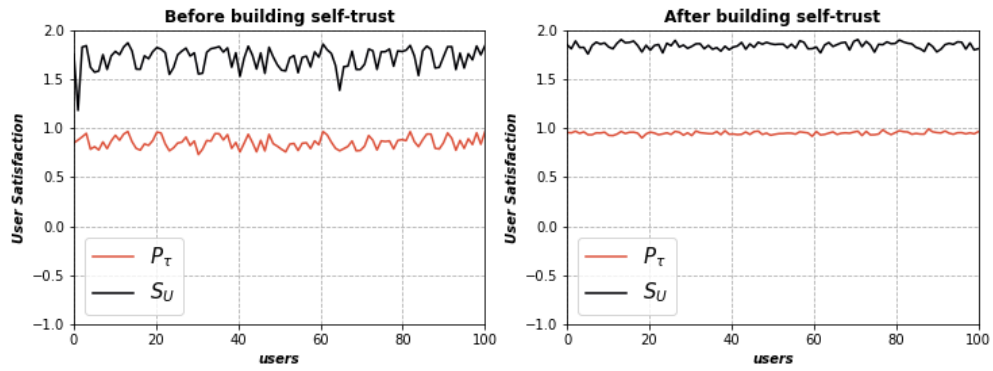
2.4 The user's satisfaction function

We have introduced a *user's satisfaction* function $S_{\mathcal{U}}$ that computes the global user's satisfaction regarding the collaboration offered by the robot; the robot aims at maximizing this function every time it interacts with a new user. $S_{\mathcal{U}}$ is the linear combination between a term P_{τ} that measures how much the user has been satisfied by the results of $\mathcal{R}$ in performing precisely the delegated task and a term $S_{\mathcal{U}plus}$ that measures how much the user has been satisfied by the additional, not explicitly required part of the plan performed by the robot in its smart collaboration. Both terms are are affected by the robot's capabilities to profile the user and to learn their corresponding trustworthiness. In particular, $\mathcal{R}$'s profiling capability is quantified by calculating how the robot has adapted the task to the real user's features that form $\mathcal{P}_{\mathcal{U}}$: the greater is this measure for each feature, the more accurate is the robot's capability to profile the user on that feature and the greater are the user's satisfaction components mentioned above. As will be clear in the results section (section 3), both components P_{τ} and $S_{\mathcal{U}plus}$ are designed so that they vary in the codomain $[0, 1]$, while $S_{\mathcal{U}}$ varies in the codomain $[-1, 2]$.

## 2.5 The experiment's phases

The experiment is structured as follows:

1. the robot implements a first trial with a population of class $\mathcal{C}_1$. During this multiple interaction, the robot decides to adopt the task at the level of help it considers appropriate to the user and the context. The phase is designed so that $\mathcal{R}$ infers the feature *category* with a low level of accuracy, while the other features of $\mathcal{P}_{\mathcal{U}}$ are inferred with an high level of accuracy;

2. the robot implements a second trial with the same population of class $\mathcal{C}_1$ exploited in the previous phase. During the trial, the robot decides to adopt the task at the level of help it considers appropriate to the user and the context. In this case $\mathcal{R}$'s capability to infer $\mathcal{P}_{\mathcal{U}}$ degrade because the features *age, category, education* are affected by a low level of accuracy (features *company* and *economic status* are still inferred with an high level of accuracy);

3. the robot starts a training phase with a new population of class $\mathcal{C}_1$, in order to learn its own level of self-trust. In this phase, $\mathcal{R}$ has the same profiling skills described at point 1. Please recall that, during the training phase, $\mathcal{R}$ adopts the task at a literal level of help;

4. the robot starts a second training phase with a new population of class $\mathcal{C}_1$, but this time its profiling skills are the same described at point 2;

5. the trials described at points 1 and 2 are repeated, but this time the robot exploits what it has learned respectively in the context described at the point 3 and 4, in order to achieve the task adoption process.

(a)
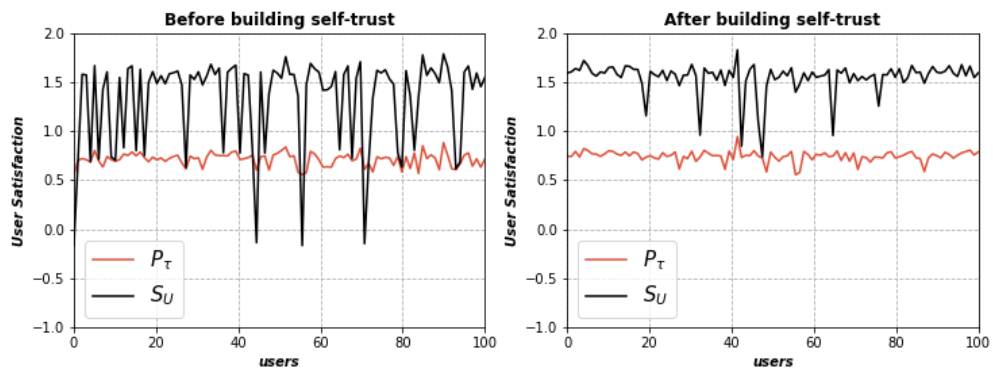
(b)
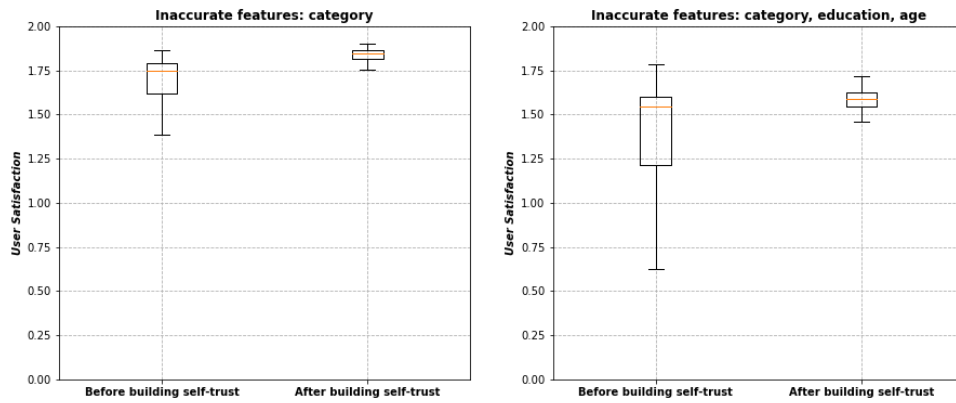
(c)

Figure 1: Figure 1a and 1b show the trend of the curves representing the user's satisfaction, obtained after each phase described in section 2.5: each plot represents the trend of the component P_τ (light red line) and the trend of S_U (dark red line) as combination of P_τ and $S_{U\,plus}$. Figure 1c shows a statistical description of the impact of the self-trust building process in the level of user's satisfaction on the robot's smart collaboration.

3 Results

In this section we present the results of the experiment designed in order to address the research purpose previously defined: demonstrate how building robot's self-trust is a precondition for providing smart and trustworthy collaboration, every time a user requires the robot's help. The plots shown in figure 1 compares the results obtained after the execution of each experiment's phase described in section 2.5.

Let's start by describing the figure 1a. This plots refer to the case when the robot's capability to recognize the feature *category* is inaccurate, while are accurate the capability to recognize the remaining features collected in $\mathcal{P}_{\mathcal{U}}$. The left plots show the distribution of P_τ and $S_{\mathcal{U}}$ obtained when $\mathcal{R}$ performs a trial with a population of class 1 and it is not yet able to evaluate the level of trustworthiness of its profiling skills. Instead, the right chart shows P_τ and $S_{\mathcal{U}}$ trends when the robot's capability are the same described in section 2.5 at point 1, but it has learned to self evaluate the trustworthiness of its own profiling skills.

Figure 1b displays the trends of the user's satisfaction function $S_{\mathcal{U}}$ and its component P_τ in case the robot performs a trial with a population of class 1 and its profiling skills are such that it cannot correctly recognize the features *age,category,education*, while it infers the user's *economic status* and *company* with an high accuracy (point 2 described in section 2.5). In particular, the left part of the figure shows the results in case the robot is not able to self evaluate the trustworthiness of its profiling skills, while the right part shows how the user's satisfaction change once the robot has learned to attribute a specific level of trustworthiness to its profiling skills.

Finally, figure 1c shows the box plots comparing the distributional characteristics of $S_{\mathcal{U}}$ before and after the robot's self-trust building process. In particular, the left box plot and the right box plot refer to the cases of the robot is capable to profile the user with the conditions described respectively at point 1 and 2 of the section 2.5.

Comparing the plots in figure 1a we observe how the robot's capability to recognize the level of trustworthiness of its profiling skills is crucial for maintaining an high level of the user's satisfaction about the robot's performance. This capability become more important when the robot decides to adopt the delegated task to a level of help different with respect the literal one. Indeed, despite the robot provides unexpected results to the user, its own capabilities to adapt these results by leveraging on the capabilities that it considers trustworthy, allows the robot to provide unexpected but suitable results, that are appropriate to the user himself/herself. The plots in figure 1a and the left box plot of figure 1c, show how the mean (and the median) value of $S_{\mathcal{U}}$ increases after the robot has learned its self-trust level; moreover, the spread and the skeweness of the $S_{\mathcal{U}}$ distribution is drastically reduced by the robot's capability to self evaluate the trustworthiness of its profiling skills. Figure 1b and the right box plot of figure 1c show the benefits of the building self-trust process on the user task performance evaluation. In this case, the increase of the median value of $S_{\mathcal{U}}$ is less evident than for the previous case analyzed, but the training phase impact remains evident on the spread and the skeweness of the distribution. This means that, also when the robot's profiling skills degrade, its capability to evaluate their trustworthiness continue to allows the robot to provide unexpected but suitable results with respect to the needs of the users. It is also relevant to underline how the effective performance of the robot's help depends on the width and variety of the database of the accessible services with respect to to the selected features. In fact, with a very low number of trustworthy features (given the low level of accuracy of three of them) the result of the adoption could be really very good only if the database contains services responding, with very high performance, to the two remaining features independently to the values of the three (degraded) features.

4 Conclusions and future works

Cooperation is one of the main social activities exploited by humans for gaining resources, in terms of goals achieved, shared knowledge and so on. The increasing intelligent technology surrounding us is becoming crucial for our own social development, and, as a consequence, the need of trusting these supporting and sophisticated tools is becoming every day more stringent. But, if on the one hand these systems are becoming more intelligent and sophisticated, on the other hand they show a strong lack in the ability to collaborate effectively with humans. Despite the complexity of the problem they can solve, they continue to have just a passive supporting role in the collaboration with humans. For being not only executive tools, these intelligent systems (i.e. robots, chatbots, autonomous cars and so on) should expose the capability to behave in a critical way with respect to the needs/goals of their interacting users. Indeed, the collaboration becomes deep and effective when a system is able to provide not declared, unexpected results but compatible with the context, the needs of the user and the capabilities of the system itself. The level of autonomy of robots or other artificial agents, it should be such that such systems can exercise a certain level of discretion in achieving the task delegated but humans. But, in order to foster trust in humans, they should behave having the capability to create a complex theory of mind of the

interlocutors and a strong capability to self assess their own capability to carry out a task, also at a different level of help than required.

In this work we have presented the first of a series of experiments draw for testing different aspects of a designed cognitive architecture. This architecture, based on consolidated theoretical principles (theory of adoption and delegation, theory of mind, theory of social adjustable autonomy, theory of trust) has the main goal to build robots that provide smart, trustworthy and transparent collaboration, every time a human requires their help. With this experiment we wanted to test the robustness of the designed architecture to rely on the robot's ability to learn the limits in interpreting the needs of its interlocutor, by measuring the trustworthiness of its predictive abilities. In fact, the architecture gives to a robot the capability to profile the user and to leverage on its profiling skills in an adaptive manner, by exploiting those skills that maximize the user's task performance evaluation; it allows the robot to reason about the mental states of the user (beliefs, goals, plans and intentions) and makes it capable to modulate its autonomy for achieving the delegated task. One of the main problems in intelligent collaboration between humans is the possibility of misunderstandings that can lead to conflicts between cooperators. We call these collaborative conflicts, as they are based on the desire to collaborate beyond what is required but in doing this errors and discrepancies can occur. Just to minimize these conflicts and increase the robot's trustworthiness, an important requirement to introduce is the the capability of the robot itself to self trust his capabilities to build a complex model of the user. The data analyzed have shown how the process to learn the trustworthiness of its own profiling skills can lead the robot to have an effective collaboration, based not only on the actions/tasks prescribed by the user, but especially on the non declared needs and goals of the user himself/herself. Our main future work will be to move the experiment in a real environment, with a real robotic platform and real users. We will exploit the humanoid robot *Nao*, widely used in HRI applications. Furthermore, we will continue to provide simple but effective experiments that allow us to investigate different aspects of the concept of intelligent and trustworthy collaboration between robots and humans, that consider robots as cognitive agents able to interact with humans as humans do when they interact with each others.

References

[BBH+13]   Olivier Boissier, Rafael H Bordini, Jomi F Hübner, Alessandro Ricci, and Andrea Santi. Multi-agent oriented programming with jacamo. *Science of Computer Programming*, 78(6):747–761, 2013.

[BER15]   Matthias Braunhofer, Mehdi Elahi, and Francesco Ricci. User personality and the new user problem in a context-aware point of interest recommender system. In *Information and communication technologies in tourism 2015*, pages 537–549. Springer, 2015.

[BH05]   Rafael H Bordini and Jomi F Hübner. Bdi agent programming in agentspeak using jason. In *International Workshop on Computational Logic in Multi-Agent Systems*, pages 143–164. Springer, 2005.

[CF98]   Cristiano Castelfranchi and Rino Falcone. Towards a theory of delegation for agent-based systems. *Robotics and Autonomous Systems*, 24(3-4):141–157, 1998.

[CF10]   Christiano Castelfranchi and Rino Falcone. *Trust theory: A socio-cognitive and computational model*, volume 18. John Wiley & Sons, 2010.

[CF20]   F. Cantucci and R. Falcone. Towards trustworthiness and transparency in social human-robot interaction. In *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, pages 1–6, 2020.

[DA16]   Sandra Devin and Rachid Alami. An implemented theory of mind to improve human-robot shared plans execution. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 319–326. IEEE, 2016.

[FC01]   Rino Falcone and Cristiano Castelfranchi. The human in the loop of a delegated agent: The theory of adjustable social autonomy. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 31(5):406–418, 2001.

[HMDAR16] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. The off-switch game. *arXiv preprint arXiv:1611.08219*, 2016.

[IAF⁺19]   Brett Israelsen, Nisar Ahmed, Eric Frew, Dale Lawrence, and Brian Argrow. Machine self-confidence in autonomous systems via meta-analysis of decision processes. In *International Conference on Applied Human Factors and Ergonomics*, pages 213–223. Springer, 2019.

[KS20]   Bing Cai Kok and Harold Soh. Trust in robots: Challenges and opportunities. *Current Robotics Reports*, pages 1–13, 2020.

[MKI19]   Marwa Hussien Mohamed, Mohamed Helmy Khafagy, and Mohamed Hasan Ibrahim. Recommender systems challenges and solutions survey. In *2019 International Conference on Innovative Trends in Computer Engineering (ITCE)*, pages 149–155. IEEE, 2019.

[Paz99]   Michael J Pazzani. A framework for collaborative, content-based and demographic filtering. *Artificial intelligence review*, 13(5):393–408, 1999.

[RG⁺95]   Anand S Rao, Michael P Georgeff, et al. Bdi agents: from theory to practice. In *ICMAS*, volume 95, pages 312–319, 1995.

[SADL18]   Hannah J Swift, Dominic Abrams, Lisbeth Drury, and Ruth A Lamont. Categorization by age. *Encyclopedia of Evolutionary Psychological Science*, 2018.

[SSTJS18]   Sarah Strohkorb Sebo, Margaret Traeger, Malte Jung, and Brian Scassellati. The ripple effects of vulnerability: The effects of a robot's vulnerable behavior on trust in human-robot teams. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 178–186, 2018.