A Cognitive Approach to Model Intelligent Collaboration in Human-Robot Interaction

Filippo Cantucci[*,†], Rino Falcone[†]

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

Abstract

While Human-Robot Interaction (HRI) and Social Robotics achieved significant results, research has not given the right attention both to the role of the analysis of cognitive processes by a robot and an extensive interpretation of intelligent collaboration, i.e robot adaptivity, trust, categorization, and so on. Daily life scenarios involving robots require a kind of collaboration oriented towards the respect and the achievement of the most relevant users goals, but also to the validation and protection of other application domains elements (e.g. other agents, available resources). To address this challenge, computational cognitive models are required, in order to integrate multiple users goals and interests within the interaction context and support robots to justify the integration process. In this theoretical paper, we propose a taxonomy, achievable although with certain constraints and occasional imprecision, of the human action intentions into three distinct classes: practical, socio-normative, and ethical. Building upon this premise, we outline the development of computational cognitive models aimed at supporting a useful, effective, acceptable and trustworthy interaction between humans and robots. While our conceptual framework is applicable across various domains, we will illustrate it with examples drawn from the domain of Cultural Heritage.

Keywords

Human-Robot Interaction, Social Robotics, Robot Adaptivity, Cognitive Systems, Trust

1. Introduction

In today's human-centric world, interactions with autonomous systems, such as robots, become increasingly ubiquitous. To foster a real sense of *dependability* and *trust* in artificial intelligent systems, it is demanding to design systems that are more *useful*, *effective*, *acceptable* and *trustworthy*. As articulated by Castelfranchi and Falcone [1], trust is not just the outcome of the frequency with which an agent produces the expected result; trust is a much more complex attitude, including a *causal attribution*, an *estimation*, an *ascription* of several internal factors that play a causal role in the behavior activation and control.

To foster trust in Human-Robot Interaction (HRI), humans should perceive the robot's behavior in mental terms, by ascribing a mind with their representation [2]. Such cognitive interpretation and narrative of the mechanisms controlling and orienting their behavior, is just the description at the macro functional level of the underlying micro-computation. The

WOA 2023: 24rd Workshop From Objects to Agents, November 6–8, Rome, Italy

[*] Corresponding author.

[†] These authors contributed equally.

filippo.cantucci@istc.cnr.it (F. Cantucci); rino.falcone@istc.cnr.it (R. Falcone)

micro-level implementation of the regulatory behavioral process is unintelligible [3] and even irrelevant at that level. Much like in psychology, humans leverage cognitive concepts like belief, reasoning, decision-making, desire, intention, and so on, to provide a functional description of how the brain operates. Humans anthropomorphise anything, including AI systems. In such *anthropomorphization* they identify other forms of *intelligence*, i.e. behavior regulation based on representations and mental/virtual problems or solutions; thus they build a broader, unified theory of such notions. This approach is evident in collaborative scenarios, like shared plans, joint action, task delegation.

In multiple everyday robotics scenarios (e.g. elderly assistance [4, 5], cultural heritage [6, 7], tourism [8] and so on), it is crucial that the robot's collaboration takes into account both the user's primary goals achievement and the validation and safeguarding of other contextual elements, including other agents, available resources, interested stakeholders. In these contexts, *intelligent collaboration* implies more *autonomy* and even initiative, because it requires more than simple execution of a prescribed action or plan. Despite some remarkable result in HRI [9], research has typically overlooked the role of a robot's cognitive processes analysis and its interpretation in terms of intelligent collaboration, including smart adoption, trust, categorization, and so on. Likewise, in the HRI or Social Robotics domains, robots should provide humans with different levels of help [10]: they should leverage their autonomy, competence and cognitive abilities to identify better or potential solutions for their goals and needs. This could be autonomously initiated by the system without the need for negotiation, discussion, or agreement [11]. How could this advanced form of collaboration be possible without a reciprocal mind ascription and reading? A robot needs to understand human's ends, higher goals, expectations, interests, in order to really help them. And human have to understand what robot do in mental terms, including its knowledge, beliefs, reasoning and problem-solving strategy. Furthermore, the robot has to explain what it did or plan to do in terms of the *reasons* for doing so: data, prediction, objectives, solutions ans so on. There is no real cooperation without minds representing minds.

An interesting approach to deal with this challenge is:

- to model the user's mental states, including their goals, beliefs, plans, and interests, as well as those of other relevant stakeholders, all aimed at various ends of their actions;
- to separate the *practical*, *social/normative* and *ethical* aspects;
- to compose these different types of mental states, establishing priorities and compatibility, by exploiting different levels of autonomy in achieving tasks;
- to explain if and how this composition occurred.

Theory of Mind (ToM) [12] is a crucial ability that a robot must possess in order to implement this type of systematic approach. This involves the robot's ability to attribute mental states to the user and exploit these mental states for decision-making.

1.1. Usefulness, Effectiveness, Acceptability and Trust in HRI

Four interaction's attributes are crucial in HRI scenarios: *usefulness*, *effectiveness*, *acceptability* and *trustworthiness*. They can be defined as follows:

- *Usefulness*: humans should exploit interaction with robots in order to achieve their own goals.

- *Effectiveness*: interaction should realize those results, in times and ways that are satisfying for the specific human subject.
- *Acceptability*: humans should recognize and agree with the role played by robots in the interaction, and they should acknowledge and accept the robot's prerogatives in assuming that role.
- *Trustworthiness*: human should *trust* the robot, in order to start the interaction and recognize its potential benefits. This entails the human perceiving the robot as possessing both *abilities* and *will/motivations* to pursue the task that will yield the aforementioned benefits during the interaction. We will expand this concept in section 1.3.

We can analyze the properties defined above in HRI, starting from the main humans needs when they interact with each other. In any collaborative scenario, every time humans delegate tasks to other humans, they hold expectations that include, at the very least:

1. *to pursue their own goals*;
2. *to consider broader interests and goals.*

In the case 1, the robot has a defined goal to pursue on behalf of the human in that specific interaction. Firstly, the robot should be able to accurately understand the human's goals and how to accomplish them. Furthermore, it should achieve these goals while considering the effectiveness and acceptability of its role. In addition, its trustworthiness must be assessed. The effectiveness will depend on the plan that the robot intends to implement to achieve the goal, and, therefore, on the boundary conditions that will be implemented. This, in turn, will influence human satisfaction with those conditions.

In the case 2, the robot should understand interests and goals beyond the specific interaction and provide support for their achievement or protection. For instance, the task explicitly delegated by the human to the robot could be a part of a more complex goal to which the robot might be involved, or it could be in contradiction with other goals or interests of the same human, which the human has not consciously evaluated, and which the robot should consider. The robot doesn't merely respond to a request; it conducts a more complex assessment of the human's intentions, beliefs, motivations, and interests. Therefore, it implements a plan capable of satisfying these more complex attitudes. In this case, conflicts can arise, leading to potentially interesting collaboration. To provide an explicit example in the field of Cultural Heritage, consider a human visitor to a technologically advanced museum who provides sensors to collect physiological parameters that can be correlated to the emotional states of its users. Now, imagine that when the visitor asks the robot to walk into the room where a famous Picasso painting is placed, the robot decides to suggest a different continuation of the tour. These suggestions could vary significantly, eliciting different reactions. For instance, the robot might suggest, based on the user's profile and an interest in the Italian Renaissance, to visit Michelangelo's room instead of seeing the Picasso exhibit, which is in a very crowded room with long access times. Alternatively, the robot might recommend a short break at the bar to address detected low blood sugar levels, or it could even call a doctor or ambulance upon detecting a dangerous abnormality in the person's heart rate. In these cases, the robot takes into consideration goals and interests belonging to the user, even if the user isn't consciously considering or recognizing them, and these may not align with the task initially assigned by the user.

1.2. Conflicts in intelligent HRI

It is evident that a form of collaboration as described above can rise conflicts between humans and robots. These conflicts can be categorized as follows:

- *Misunderstandings*: the human may not understand the meaning of the robot's behavior unless explicitly stated. This can occur when the robot decides to pursue goals different from those explicitly articulated by the human, even if these alternative goals align with the human's own objectives or areas of interest;
- *Doubts*: the user doubts that the robot is truly motivated by tutorial goals for the human's benefit or if it is pursuing its own goals or non-cooperative related to other subjects.
- *Susceptibility*: the human cannot consider the behavior acceptable and may disagree;
- *Refusals*: the human cannot consider the robot entitled to make such a proposal in defense of goals and interests that do not fall within robot's prerogatives.

The previously described approach to intelligent help for the achievement of goals and interests beyond the explicit request require addressing, at least, doubts or misunderstandings. Achieving this requires supporting collaboration through communicative actions by robots capable of justifying their assumed roles. *The explanation* of the reasoning process is grounded in the attributions the robot has made regarding the user's goals and interests, the value of priorities identified at different contextual levels, and the chosen satisfaction model for assistance. This approach should make the rationale behind the algorithms, which aim to find the optimal solution for user satisfaction, clear. Utilizing a blend of graphical and spoken mechanisms for presenting these explanatory schemes can enhance the clarity of the collaboration model for the user.

1.3. Trust in HRI

As mentioned above, a crucial aspect of human-robot interaction concerns trust. Robots should be capable of both being trustworthy and trusting other agents, whether they are human or artificial [13]. Analyzing the complex concept of trust in the context of robot collaboration, as previously described, is a challenging task. One possible approach to tackle this challenge is by adopting the socio-cognitive trust model developed in [1]. We introduce the concept of modelling trust/trustworthiness [14] in the following ways:

- *Humans' trust in collaborative robots*: we consider how robots, able to carry out behaviors with differentiated tutorial levels of intelligent help, are considered trustworthy by humans.
- *Robots able to trust*: we develop a robot's trust model based on different sources of trust (direct experience, reputation, types of reasoning, etc.) and then evaluate different performances of these behaviors.

As we said, it is possible to articulate goals and interests of the user not only on the basis of practical needs but also taking into account social, normative, ethical needs. In this sense, it is possible to plan a trust-based interaction model for the robot in order to analyse these different needs separately.

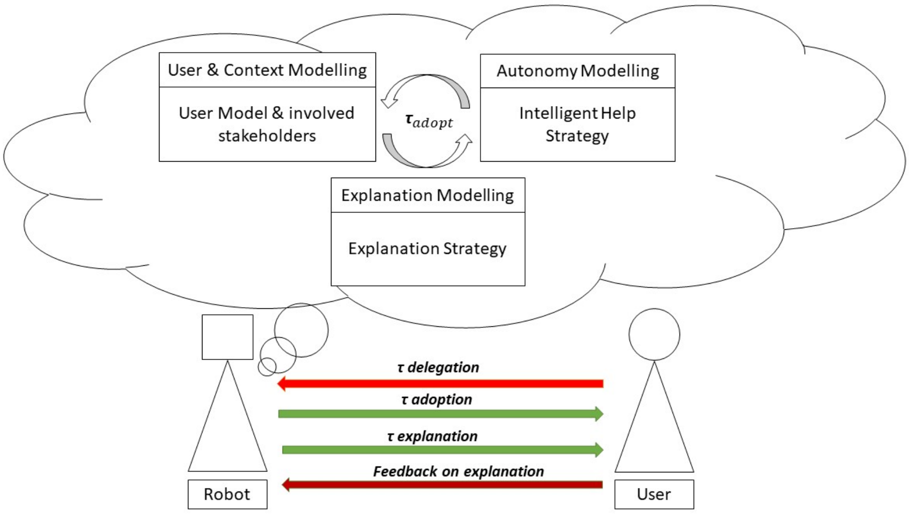

Figure 1: The explanation represents a process that contributes to the definition of the adoption process.

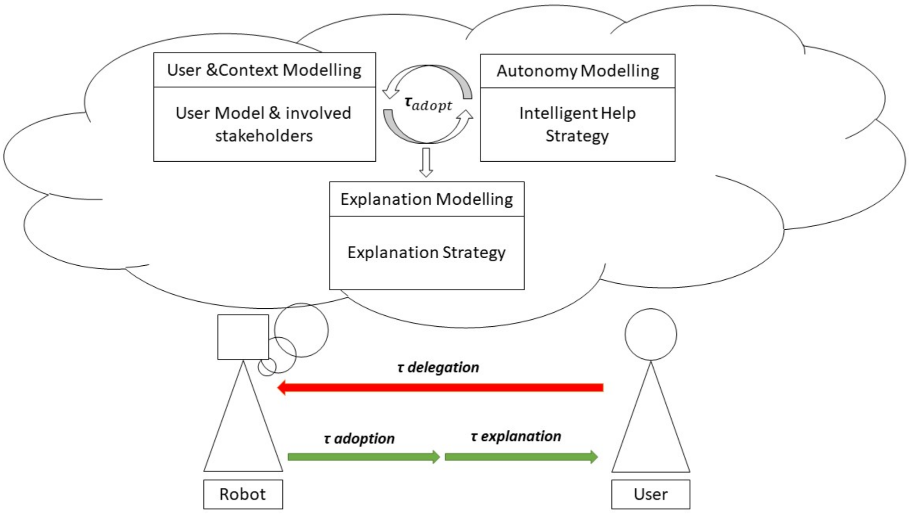

Figure 2: The explanation represents a tool for understanding the adoption, if the user receives the final results before or in conjunction with the explanation process.

2. Basic elements for implementing Intelligent HRI

What was discussed in Section 1 can only become possible through the definition of a cognitive architecture [15, 16] that supports a robot's decision-making system, enabling it to acquire complex cognitive capabilities. In this paper, we introduce a conceptual framework for identifying the fundamental elements for building computational cognitive models that support

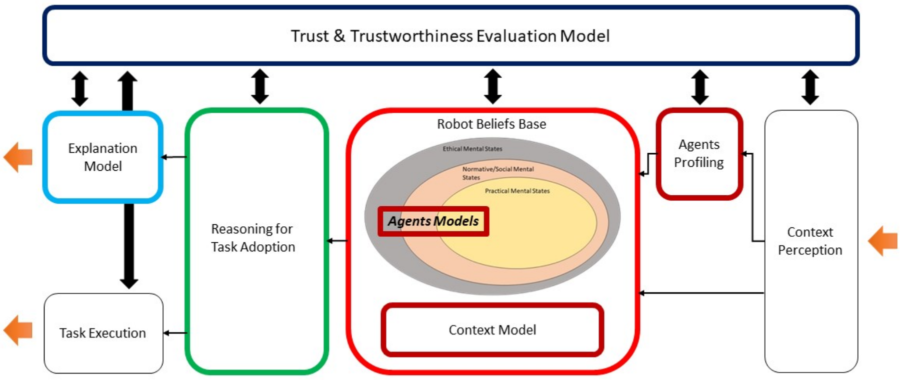

Figure 3: Performing adaptive and effective reasoning in a dynamic interaction with other cognitive agents, typically humans, on the basis of several context factors such the practical, social or ethical goals/needs attributed to the human users involved in the interaction. Modules in red model the interaction context described in section 2.1; modules in green model the task adoption strategy described in 2.2; modules in green model the task adoption strategy described in 2.2; modules in blue model the task explanation described in 2.3; modules in purple model the trust evaluation described in 2.4.

a useful, effective, acceptable and trustworthy interaction between humans and robots. The design model follows the following lines:

- *Representing Interaction*: this involves representing the interaction through a set of mental states of the robot, as well as attributing these mental states to other participants in the interaction, such as the user and other stakeholders.
- *Intelligent Adoption Strategy*: the model incorporates an intelligent adoption strategy that enables the robot to adapt its level of collaborative autonomy based on the representation of mental states mentioned above, including those that reflect the user's needs. This adjustment is made possible through an adaptable autonomy model, offering various levels of delegated task adoption.
- *Explanation Strategy*: the model includes a strategy for explaining the adoption process. This explanation aims to facilitate the end user's understanding of the reasons that led the system to adopt the delegated task. The explanation can be an integral part of the adoption process (as depicted in Figure 1) or provided as a separate tool for understanding the adoption, particularly if the user receives the final results before or concurrently with the explanation process (as illustrated in Figure 2).
- *Trust & Trustworthiness evaluation*: please refer to section 2.4 for a comprehensive discussion.

The Figure 3 illustrates the fundamental components for designing computational cognitive models that enable a task-oriented intelligent artificial agent, such as a robot, to engage in adaptive and effective reasoning during dynamic interactions with other cognitive agents, typically humans. These interactions depend on various contextual factors, including practical, social, or ethical goals and needs attributed to the human users involved in the interaction. In the following sections, we will describe the key modules depicted in the figure.

2.1. Modelling the Interaction Context

The primary objective of these modules is to provide an intelligent robot with a knowledge representation suitable for the specific nature of its collaboration. This enables the agent to formulate *plans* for actions based on: i) its knowledge of the ever-changing and dynamic world; ii) the user it's collaborating with; and iii) the normative or ethical constraints within the interaction environment, which are represented as mental states attributed to the user or other agents involved in the collaboration. To align the task adoption process with the user's needs, the agent must profile the user at different levels. For instance, it needs to collect information pertaining to the classes of mental states we've previously defined (practical, social, ethical mental states). The outcome is an ontological framework that represents the artificial system's knowledge while considering the relationships between goals, actions, and rules/norms/ethical principles applicable to actions, as depicted in Figure 4.

To illustrate these relationships between practical/socio-normative and ethical states of the world, consider a museum user who finds themselves in room R_1 and requests more fresh air due to feeling overheated. Suppose the robot has a plan $\pi_1 = openWindowIn(R_1)$ in its action repertoire that results in providing fresh air to room R_1. Let's imagine that π_1 involves opening a window in room R_1. From a practical perspective, the robot can execute this plan if all the practical preconditions of π_1 are met, for example having the strength and ability to open the window and ensuring unobstructed access to it. However, from a socio-normative perspective, the robot should consider additional constraints. For instance, it should verify if there are any rules prohibiting the opening of windows in the museum. Even in the absence of such a prohibition, the robot should assess whether opening the window might disturb other visitors. Furthermore, from an ethical standpoint, the robot should evaluate the potential risks. For instance, an open window with a very low sill might endanger the lives of small visitors. This would not only violate a formal rule but also compromise an ethical principle of safeguarding the lives of other humans. In this example, the constraints are more objective than subjective to the user. However, the robot, armed with an accurate model of the user, could also take into account specific socio-normative and ethical constraints. It could compare these with other constraints, such as general rules, shared ethical principles, as well as specific regulations and social constraints established by museum managers and curators.

In general, each robot accesses a plan library Π that plays a crucial role in the robot's decision-making process and it is defined as follows:

$$\Pi = \Pi^a \cup \Pi^d \tag{1}$$

where Π^a indicated abstract plans (i.e., walk, drive, run are a type of move), while Π^d is the set of plans that can be decomposed in sub-plans (i.e., travel consists of leaving, moving and arriving). We consider the plan library as unique and common to all agents, both robots and humans. Each plan is a tuple $\pi =< p, \gamma, \lambda, b >$ (with $\pi \in \Pi$) that has a body b (course of

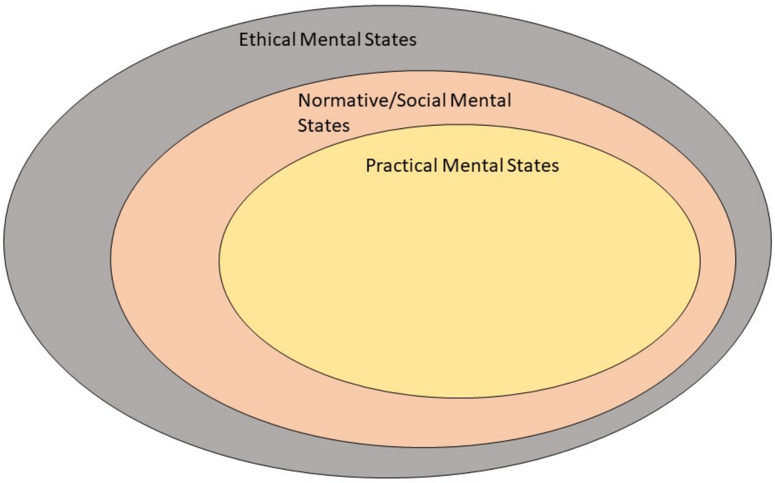

Figure 4: Venn's diagram representing (hierarchically) the user's mental states referred to different domains: Practical, Socio-Normative, and Ethical.

actions α_i to execute or sub-goals g_i to achieve), preconditions γ, constraints λ and states of the world p it will realize if correctly applied. Please refer to [17, 18] for a much more complete meaning of plans. If the preconditions and constraints refer only to the practical world, then we remain in the practical domain. On the other hand, in the case in which the preconditions and constraints, but also the results, involve formal norms of the community or influence the sphere of other subjects, then we are in the socio-normative domain. Finally, if they involve the ethical principles of the subject to whom the robot's collaboration is addressed, then we are in the ethical domain.

2.2. Modelling the Interaction/Task Adoption strategy

As previously mentioned, the adoption strategy is executed by the robot through actions in the real world. The system maintains a repertoire of actions and a plan library that organizes these actions in a coherent manner to achieve specific states of the world, such as goals. The syntax used aligns with the typical format of practical reasoning plans [19, 18]. A crucial aspect in ensuring effective collaboration is the creation of a plan library capable of distinguishing between three defined context: practical, socio-normative, ethical. When the robot adopts a delegated task, it assesses whether to consider only the practical utility or to include socio-normative or ethical considerations. It is important to emphasize that even when the robot chooses to focus solely on the practical context, it can still provide intelligent help rather than a purely *literal* response to the delegation. In fact, it can take into consideration other user's *active goals*, which are those goals that the user could reasonably have decided to achieve (through the delegated goal or otherwise) and that the robot is able to foresee, through the modeling of the user [20, 7]. In such cases we can refer to this help as *critical help* or *overhelp*.

Given a *practical* task τ delegated by the user, the robot can implement different forms of adoption, which we can categorise into:

- *Practical Help*;
- *Socio/normative Help*;
- *Ethical Help*.

The basic assumption is that when practical help is required, the robot can adopt that help by considering to consider only the constraints imposed by the practical world. However, it may also evaluate and potentially question socio-normative constraints related to the specific user, and even take ethical constraints into account, all while considering the needs of that user. Moreover, each category of help ensures that the intelligent robot can adopt the task in a literal, critical way or taking into account the user's active goals that the user itself has not explicitly requested to achieve [21]. The choice of approach can vary depending on the classification of mental states within the intelligent system.

To illustrate how a robot can utilize socio-normative and ethical considerations, let's assume again that we are with a human visitor to our technologically advanced museum who provides sensors and other information to collect physiological parameters. Imagine that the user has requested the robot to plan a museum visit, including a newly decorated stone room. The literal help on the practical task would lead the robot to define the tour with the required special room included at the end of the visit. However, by analyzing the museum's room occupancy and foreseeing a potential long queue when the user reaches the decorated stone room, the robot detects a risk of the user missing the visit due to museum closing hours. Understanding that the user belongs to the category of individuals requiring special assistance (e.g., in a wheelchair), the robot leverages socio-normative context to issue a priority pass (literal help in the social-normative context). This allows the user to visit the decorated stone room at the end, prioritizing accessibility. However, this act of prioritization may raise ethical concerns since it could affect other visitors' experiences. Consequently, the robot, mindful of the ethical values, revises the user's tour (Literal help in the ethical context). It changes the order of room visits, ensuring fairness and avoiding the exclusion of other visitors from the decorated stone room. In this scenario, the robot doesn't just solve the practical problem of planning a tour but also strives to accommodate socio-normative constraints and ethical considerations that may arise, thus enhancing the visitor's experience and ensuring fairness for all museum-goers.

2.3. Modelling Explanation

The task adoption process that leads an agent to define a final intention (goal to adopt) to carry out, is based on the set of beliefs, goals, plans (mental states) it exploited during the entire collaboration with the human user. As argued in [22], the process of intention formation and intentional action execution is strictly based on specific sets of beliefs. There is a belief-based model of goal processing that leads an agent to create the intention it commit to carry out, as final outcome of goal-driven action generation. Based on this assumption and on the model of belief-based goal processing proposed by [23], we design an explanation model, on the basis of two conceptual layers, in order to foster behavior interpretability and explaination.

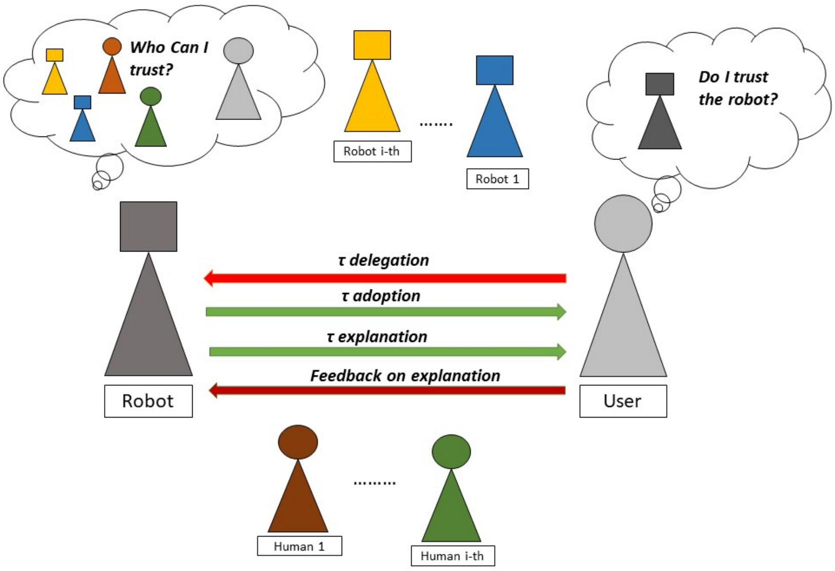

Figure 5: A robot can be trusted or trust

The first layer (*categorization*) provides the robot with the capability to categorize beliefs and goals involved in the intention creation process. In particular, each belief, which can be by definition practical, social or ethical, is classified on the basis of their support to the decision to determine a progression of the goal that will become the intention. Indeed, the intention creation process is founded on multiple progressive *goal states* and *categories of supporting beliefs* (see [23]). In addition to the beliefs, the capability to have a theory of mind of the user, allows the robot to reason about those goals explicitly delegated by the user and those ones that can be attributed to the user and that represent state of affairs that she could be interested to achieve or that has already planned to achieve in the future. We have already defined this kind of goals as active goals. At this stage, the robot owns a *mental map* of a set of beliefs (adequately classified) and goals involved in the reasoning process that led to choose a specific intention to adopt. The second conceptual layer (*explaination generation*) gives to the agent the capability to work on the mental states (beliefs, goals) produced in the categorization layer and that represent explanatory information (*agent's reasons*), in order to generate different, contextual, modalities of behavior explanation [24]. The overall explainability process (behavior interpretability due to the categorization process and behavior explaination due to the generated explaination) provided by the agent to the human user, is evaluated by her, on the basis of different dimensions [24].

2.4. Modelling Trust

The goal of this module is to endow the robot with the complex mental attitude of trust, related to the kind of collaboration provided by the designed robot (see Figure 5). As Castelfranchi and

Falcone formalize [1], trust is the mental counterpart of delegation. This means that robot has to be able to achieve user's goals or needs, but it has also able to delegates tasks, on the basis of an evaluation of other agents (artificial or humans) potentially involved in the interaction. Because of that, On the one hand, the computational model should give a robot the ability to estimate its own *degree of trust* in other agents, human or artificial, and on the basis of this estimate, decide whether and to whom to delegate specific tasks. On the other hand, the cognitive model should give a robot the ability to estimate its own level of trustworthiness, starting from different sources of trust linked to the user [13], in order to evaluate the best behavior to adopt.

3. Final Remarks

This work is expected to contribute, on a theoretical plan, to the field of social cognition and advanced cognitive robotics by focusing on the collaborative processes between humans and artificial intelligent systems. As far as the scientific community is concerned, the advancement of knowledge in the field of HRI and HMI interaction, in particular intelligent collaboration, represents an increasingly studied research frontier due to the growing diffusion of automatic and intelligent artificial systems in the life of every day and the enormous limits compared to an interaction with these tools that is actually natural and compatible with the cognitive attitudes of humans. This type of approach to human-robot interaction can be effective in many domains in which robots are involved. As has been demonstrated by field experiments, the use of intelligent systems capable of supporting an interaction based on analysis of user needs has proven useful in the museum sector. A robot that is able to adapt an exhibition to the user's needs can have a much more positive impact than a robot that does exactly what the user asks of it.

This type of approach, perhaps taking into consideration the gravity of the task to be carried out, can also be adopted in other scenarios, such as elderly assistance or even industry. The robotic industries themselves, increasingly interested in making their platforms permeable to an interaction as natural as possible with humans, will be involved in this enterprise. As it is known, several application areas, Industry 4.0 and smart manufacturing are increasingly characterized by the intensive use of robotic solutions that collaborate with workers to perform various tasks under their guidance. In some cases, workers interact with robots in critical environments, so commands should be entered to support accuracy and safety. Recent technological advances in social robotics are enabling new types of human-machine interactions that incorporate contextual data. It becomes crucial push research by investigating robotic solutions in which human-robot interaction is provided indirectly by the robots through real-time input. The robot can plan subsequent actions, follow a current routine, switch to an alternative routine, or trigger an alarm.

Finally, it is important to underlie that this kind of approach fits within the conceptual framework of Human-Centered Artificial Intelligence/Robotics. This means that an ethical perspective is adopted that empowers the development of an artificial intelligence always at the service of the person. In fact, if the benefits and advantages deriving from robotization are unquestionable, an ethical view based on the primacy of the human being through the full recognition of human dignity is essential.

Acknowledgments

This work has been supported by the project FAIR - Future Artificial Intelligence Research (MIUR-PNRR). We would like to thank Cristiano Castelfranchi, Valeria Seidita and Antonio Chella for the precious suggestions and discussions on this work.

References

[1] C. Castelfranchi, R. Falcone, Trust theory: A socio-cognitive and computational model, volume 18, John Wiley & Sons, 2010.

[2] C. Castelfranchi, Ascribing minds, Cognitive processing 13 (2012) 415–425.

[3] R. V. Yampolskiy, Unexplainability and incomprehensibility of ai, Journal of Artificial Intelligence and Consciousness 7 (2020) 277–291.

[4] G. D'Onofrio, D. Sancarlo, M. Raciti, D. Reforgiato, A. Mangiacotti, A. Russo, F. Ricciardi, A. Vitanza, F. Cantucci, V. Presutti, et al., Mario project: experimentation in the hospital setting, in: Ambient Assisted Living: Italian Forum 2017 8, Springer, 2019, pp. 289–303.

[5] M. Kyrarini, F. Lygerakis, A. Rajavenkatanarayanan, C. Sevastopoulos, H. R. Nambiappan, K. K. Chaitanya, A. R. Babu, J. Mathew, F. Makedon, A survey of robots in healthcare, Technologies 9 (2021) 8.

[6] F. Cantucci, R. Falcone, Autonomous critical help by a robotic assistant in the field of cultural heritage: A new challenge for evolving human-robot interaction, Multimodal Technologies and Interaction 6 (2022) 69.

[7] F. Cantucci, R. Falcone, Collaborative autonomy: Human–robot interaction to the test of intelligent help, Electronics 11 (2022) 3065.

[8] S. Ivanov, C. Webster, K. Berezina, Robotics in tourism and hospitality, Handbook of e-Tourism (2020) 1–27.

[9] D. Conti, C. Cirasa, S. Di Nuovo, A. Di Nuovo, et al., "robot, tell me a tale!": a social robot as tool for teachers in kindergarten, Interaction Studies 21 (2020) 220–242.

[10] C. Castelfranchi, R. Falcone, Towards a theory of delegation for agent-based systems, Robotics and Autonomous Systems 24 (1998) 141–157.

[11] R. Falcone, C. 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 (2001) 406–418.

[12] N. Gurney, D. V. Pynadath, Robots with theory of mind for humans: A survey, in: 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), IEEE, 2022, pp. 993–1000.

[13] R. Falcone, A. Sapienza, F. Cantucci, C. Castelfranchi, To be trustworthy and to trust: The new frontier of intelligent systems, Handbook of Human-Machine Systems (2023) 213–223.

[14] A. Sapienza, F. Cantucci, R. Falcone, Modeling interaction in human–machine systems: A trust and trustworthiness approach, Automation 3 (2022) 242–257.

[15] P. Ye, T. Wang, F.-Y. Wang, A survey of cognitive architectures in the past 20 years, IEEE transactions on cybernetics 48 (2018) 3280–3290.

[16] A. Lieto, Cognitive design for artificial minds, Routledge, 2021.

[17] M. E. Pollack, The uses of plans, Artificial Intelligence 57 (1992) 43–68.

[18] M. E. Pollack, Plans as complex mental attitudes, Intentions in communication 77 (1990) 104.

[19] M. E. Bratman, D. J. Israel, M. E. Pollack, Plans and resource-bounded practical reasoning, Computational intelligence 4 (1988) 349–355.

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

[21] R. Falcone, C. 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 (2001) 406–418.

[22] C. Castelfranchi, Intentions in the light of goals, Topoi 33 (2014) 103–116.

[23] C. Castelfranchi, F. Paglieri, The role of beliefs in goal dynamics: Prolegomena to a constructive theory of intentions, Synthese 155 (2007) 237–263.

[24] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38.