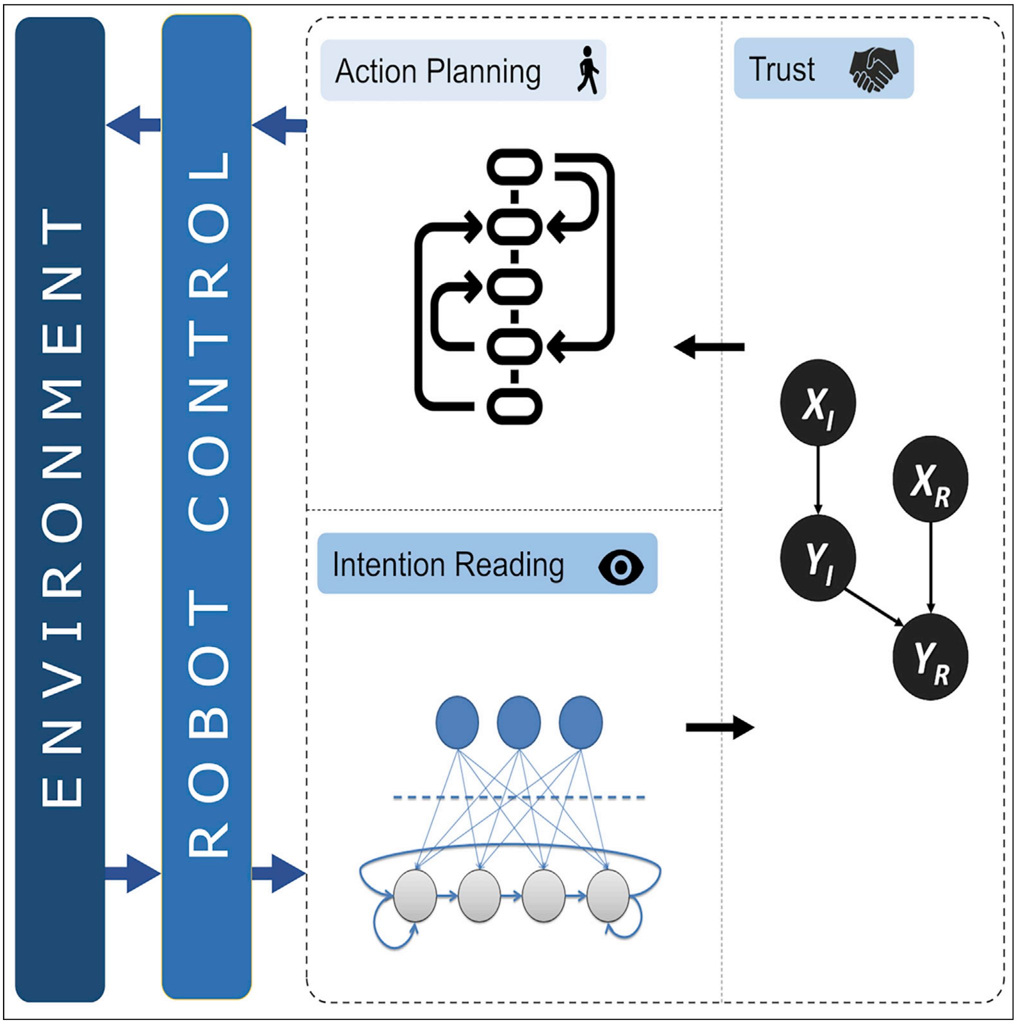
iScience

CellPress
OPEN ACCESS

Article

The collaborative mind: intention reading and trust in human-robot interaction



Samuele Vinanzi,
Angelo Cangelosi,
Christian Goerick

samuele.vinanzi@manchester.ac.uk

HIGHLIGHTS

Collaborative intelligence can be built through intention and trust estimation skills

Intention reading enables the collaboration, giving clues on the pursued goals

Trust estimation allows an agent to perform informed decision-making

Trust-aware collaborations perform better when dealing with unskilled partners

Vinanzi et al., iScience *24*, 102130
February 19, 2021 © 2021 The Author(s).
https://doi.org/10.1016/j.isci.2021.102130

iScience

CellPress
OPEN ACCESS

Article

The collaborative mind: intention reading and trust in human-robot interaction

Samuele Vinanzi,[1,4,*] Angelo Cangelosi,[1,2] and Christian Goerick[3]

SUMMARY

Robots are likely to become important social actors in our future and so require more human-like ways of assisting us. We state that collaboration between humans and robots is fostered by two cognitive skills: intention reading and trust. An agent possessing these abilities would be able to infer the non-verbal intentions of others and to evaluate how likely they are to achieve their goals, jointly understanding what kind and which degree of collaboration they require. For this reason, we propose a developmental artificial cognitive architecture that integrates unsupervised machine learning and probabilistic models to imbue a humanoid robot with intention reading and trusting capabilities. Our experimental results show that the synergistic implementation of these cognitive skills enable the robot to cooperate in a meaningful way, with the intention reading model allowing a correct goal prediction and with the trust component enhancing the likelihood of a positive outcome for the task.

INTRODUCTION

Human beings are social creatures held together by communal bonds and organized into complex social structures. This tendency to aggregation and to work as part of groups is not to be dismissed as a quirk but rather constitutes an important characteristic that has been proved being at least partially hardwired in our genes (Ebstein et al., 2010). The ability to collaborate with others to achieve common goals has been one of the key factors for our success as a species.

Researchers in the social sciences agree to distinguish collaboration from cooperation, as they represent two different types of interaction (Roschelle and Teasley, 1995). In particular, we refer to "cooperation" when the involved parties work toward a shared goal by solving sub-tasks individually and then assembling their partial results. In contrast, "collaboration" refers to the act of dividing the task among the participants, who then engage in a mutual, coordinated effort to solve the problem together. Given these definitions, the main difference between cooperation and collaboration is that the latter implies a deeper level of interaction, shared understanding, and coordination (Dillenbourg, 1999).

A body of scientific evidence points toward the early development of collaborative behaviors in human infants: the latter are, in fact, able to engage in coordinate actions as early as their first birthday. This ability continues to evolve through time and by experience, in parallel to their cognitive development, and by the 30th month of age, they become able to perform complementary actions (Henderson and Woodward, 2011).

Our hypothesis on collaborative intelligence stems from two statements. Bauer et al. (2008) break the collaboration process in a series of sequential tasks, namely perception, intention estimation, planning, and joint action. In other words, before an agent can collaborate with another, there is the need of recognizing the pursued goal and to select appropriate actions to maximize the chances of a successful outcome. Groom and Nass (2007) declare that trust is an essential component to successfully perform joint activities with common tasks. From these premises, we state that the two cognitive skills essential for successful collaboration are "intention reading" and "trust".

We refer to intention reading as the ability to understand the goals of other agents based on the observation of their physical cues, for example, body posture, movements, and gaze direction. Generally speaking, humans do not perceive biological motion as meaningless trajectories through space but instead are able

[1] Cognitive Robotics Lab, The University of Manchester, Manchester M13 9PL, UK

[2] AIST-AIRC, Tokyo, Japan

[3] Honda Research Institute Europe GmbH, Offenbach am Main 63073, Germany

[4] Lead contact

*Correspondence: samuele.vinanzi@manchester.ac.uk

https://doi.org/10.1016/j.isci.2021.102130

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

to view it in relation to an end objective (Malle et al., 2001). The cognitive process of estimating the intention is performed by dividing the observed continuous stream of actions in discrete intervals which are then individually decoded (Baldwin and Baird, 2001). By giving us the ability to understand what is happening around us, this ability lays the foundation of social awareness (Woodward et al., 2009), allowing us to reason about the behavior of other agents in our environment and acting accordingly.

Trust shares with intention reading the same importance in scaffolding our social abilities, as it affects every interaction we experience. Mayer et al. (1995) define it as the willingness of the trustor to rely on the actions of the trustee despite the former not having any control of the latter. The ability to correctly direct our trust has deep consequences on the success of our relationships, in our personal safety (Das and Teng, 2004) and in team cooperation (Jones and George, 1998).

Both these cognitive skills are not innate in humans, meaning that newborns do not automatically possess them. Instead, human phylogeny has provided each individual the tools to develop them in the scope of one's personal ontogeny, meaning that these traits will gradually arise during childhood and will refine themselves through social interactions and experiences, until reaching their full maturity. In particular, intention reading is facilitated in human beings by the mirror neuron system present in their brain (Rizzolatti and Craighero, 2004): a collection of neurons which activate both when the individual executes an action or when it observes a similar action being performed by someone else. By mapping the visual perception with the organism's own motor representation, this neurological system enables action understanding and imitation learning (Gallese and Goldman, 1998). This system is tuned by epigenetic processes during post-natal development (Ferrari et al., 2013), so it is correct to say that intention reading is perfected through experience; this is also confirmed by the fact that children are initially able to recognize biological motion, with time they start associating social cues such as biological motion and eye gaze (Tomasello et al., 2005) to goals and finally manage to understand the choice of plans (Woodward et al., 2009). In contrast, the developmental evolution of trust is still under debate. Erikson (1993) has theorized the stages of psychological development, the first of which is known as the ''trust vs mistrust'' stage that occurs around the second year of age: during this phase, the child's propensity to trust is directly influenced by the quality of cares he or she receives. This happens because infants depend entirely on their caregivers for sustenance, so if their needs are regularly satisfied, they will learn that the world is a secure and trustable place, or vice versa.

Both of these cognitive traits depend on a third one: theory of mind (ToM), the ability to understand that other beings around us possess different sets of mental states, such as beliefs, goals, and desires (Vanderbilt et al., 2011). Mastery of this capacity is a fundamental requirement for both the collaborative skills we are analyzing. In particular, intention reading can be performed only if it is possible to determine which desires are driving the actions of another agent, and trust can be estimated only if it is possible to compare beliefs and motivations to verify their alignment with one's owns (Premack and Woodruff, 1978). This dependency is emphasized by the fact that both these skills fully mature around the fifth year of age, which is also the same age at which ToM fully develops (Tomasello et al., 2005; Vanderbilt et al., 2011; Wellman et al., 2001).

Given the importance of collaborative behavior for humans, it seems natural to transpose its value to artificial agents, in particular to social robots which are expected to act in human-shaped environments interacting with us on a daily base. In particular, if we aim at designing robots able to blend themselves in our present and future societies, a strict requirement for them will be to adapt to our social expectations and fit in our natural environments. In other words, in a future where interactions between humans and robots will be more common, we do not want to robotize people, but we hope to make the minds of these mechanical companions a little more human. For this purpose, collaborative intelligence may be one of the most important skills for these agents to possess.

Collaborative intelligence, under a technical perspective, can be defined as a multi-agent system where each agent has the autonomy to contribute to a problem-solving network (Gill, 2012). For the purpose of this paper, we are interested in considering the special case of two agents, one human and a robot, which are collaborating to complete some task. In this work, we intend to expand the general collaboration architecture for cognitive robotics provided by Bauer et al. (2008) adding trust estimation between the intention reading and the action planning steps. Our proposed architecture is shown in Figure 1.

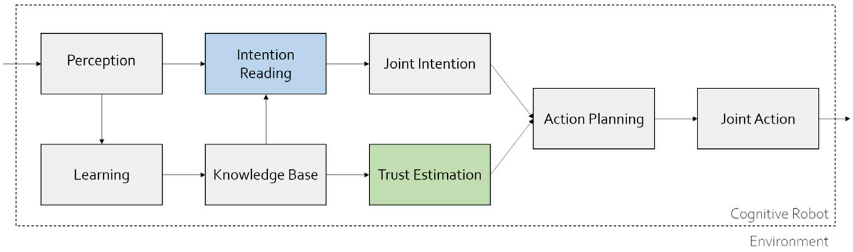


Figure 1. Overview of the mechanisms leading to joint action
Expanded from Bauer et al. (2008) through the addition of trust estimation.

The scientific community has been investigating computational models for artificial intention reading for many years, as this is an important skill for collaborative machines (Vinanzi et al., 2019a). Some techniques seem to be more common than others, in particular hidden Markov models (Kelley et al., 2008) and Bayesian networks (BNs) (Dindo et al., 2015) seem to have gained a large consensus, as well as a wide range of machine learning methods such as neural networks (Singh et al., 2016) and support vector machines (Manzi et al., 2017). Hybrid approaches have also been investigated, for example, Granada et al. (1995) used a neural architecture to extract low-level features from camera images which are then used in a probabilistic plan recognizer. The use of embodied agents such as robots for the exploration of intention reading capabilities is promoted by Sciutti et al. (2015), who underline the importance of sharing the same action space with the human partner. Robots have in fact been successfully used to investigate intention understanding and sharing in turn-based games that possess a strong learning-by-demonstration aspect (Dominey and Warneken, 2011; Jansen and Belpaeme, 2006).

Trust has also been extensively researched in the context of human-robot interaction (HRI), the main reason being that the quality of the interaction is usually shaped by how trustworthy the robot appears to the human. This means that even a perfect machine will not be able to perform at its fullest if the human partner is not willing to trust its decisions and actions. This problem has generated a branch of research focused on determining which behavioral and esthetic elements of a robot can influence its perception from the people who interact with it, in other words there is a vast literature of human-centered trust in HRI (Floyd et al., 2014; Zanatto, 2019). Here, we propose that the opposite, i.e., the trustworthiness of a human estimated by a robot, is also fundamental during a collaborative activity: whereas a robot can fail, so can a person, and it is important to keep this in mind when performing decisions that will try to optimize the achievement of the shared goal. Unfortunately, literature is scarce for what concerns this kind of robot-centered trust. Patacchiola and Cangelosi (2016) proposed a probabilistic model which unifies trust and ToM to be used in a simulation of Vanderbilt's experiment about children's trust willingness (Vanderbilt et al., 2011). This model has been subsequently expanded into a cognitive architecture for a humanoid robot (Vinanzi et al., 2019b) enhanced with an episodic memory system. The latter is a subcategory of the long-term declarative memory that stores memories about temporally dated episodes or events and temporal-spatial relations among them (Tulving, 1972). This feature is relevant because the positive influence of one's personal history on the cognitive capabilities has been proven other than for the biological brain also for artificial agents (De Castro and Gudwin, 2010; Jockel et al., 2008). Episodic memory is also the key to reproduce the ''trust vs mistrust'' stage theorized by Erikson (1993) in a developmental cognitive system.

In this paper, we present an integration of our previous studies on artificial intention reading (Vinanzi et al., 2019a, 2020) and trust estimation (Vinanzi et al., 2019b) to create a collaborative intelligent embodied agent able to direct its efforts in providing assistance in a shared activity with a human partner. Through the use of this computational model, we aim at demonstrating the positive influence of trust on the synergistic efforts of the two agents. Given this premise, our main contribution comes in the form of the novel cognitive artificial architecture for human-robot collaboration shown in Figure 2, capable to perform both intention reading and trust estimations on human partners. To achieve this, we have made use of a set of state-of-the-art techniques ranging from unsupervised machine learning methodologies to probabilistic modeling. We have validated this architecture through a set of simulated HRI experiments involving several humans and a robot collaborating in a block placing game. The results we collected demonstrate that the pairing of these two cognitive skills can greatly enhance the outcome of the joint action by providing the robot with some decision-making parameters that are used to fine-tune the assistive behavior.

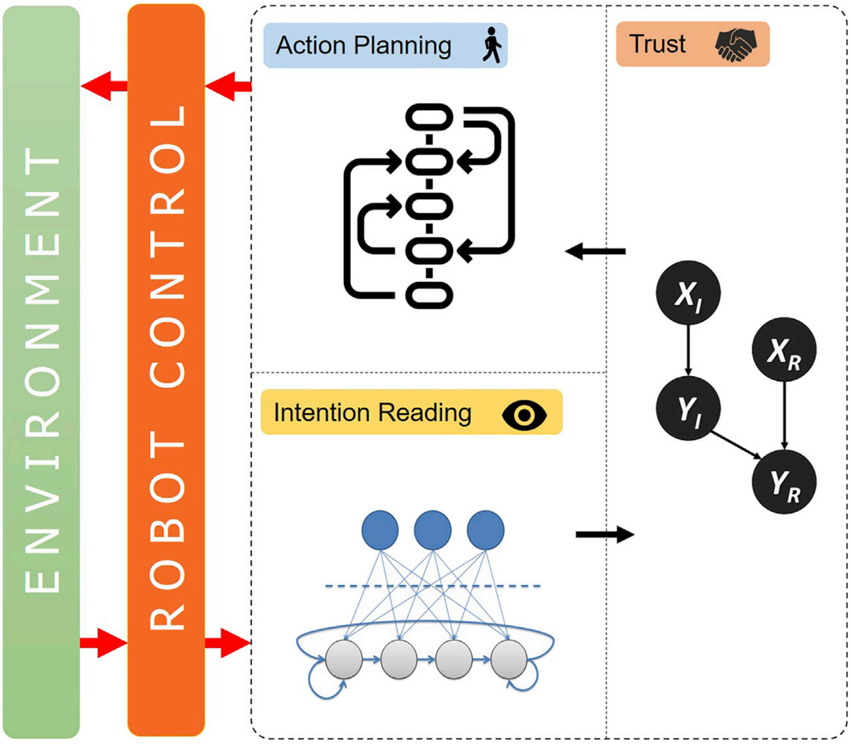


Figure 2. The proposed artificial cognitive architecture which integrates intention reading and trust mechanisms for the purpose of collaborative intelligence

Please refer to the Transparent methods section of the Supplemental information for the detail of each component.

RESULTS

Experiments

Many of the considerations made throughout this section refer directly to the methodology involved in this line of research. For this reason, we invite the reader to consult the Transparent methods section of the Supplemental information to gain a better insight on the experiments that are described below.

Having already validated the performance of our intention reading (Vinanzi et al., 2020) and trust (Vinanzi et al., 2019b) models in our previous publications, the aim of our current experiment is to verify our hypothesis on the positive influence of trust mechanisms on the overall collaborative performance. For this reason, we are going to use the same experimental setup of our previous investigation on robotic mind reading and compare the results achieved from our new, integrated architecture (referred as trust architecture, or TA) with the baseline obtained from our previous intention reading model (Vinanzi et al., 2020) which we will hereafter be referring to as the no trust architecture, or NTA.

The experimental setup is shown in Figure 3A. A Sawyer robot and a human are facing each other on the two sides of a table. Four different colored blocks are positioned on the corners of the playing area; anticlockwise from the top left they are blue (B), orange (O), red (R), and green (G). The central area of the table is denoted as the building space.

The aim of the game is to use the 4 available blocks to form a line, following a simple rule only known by the demonstrator: the blocks must be chosen one by one from a different side of the table (left or right). The 8 legal combinations of blocks are reported in Figure 3B and each of them forms a goal for our intention reading purposes. During the training phase, the human will demonstrate each goal once and the robot will learn to associate the demonstrator's body posture and eye gaze direction to their intentions. Additionally, the robot will always naively trust its teacher, while the beliefs regarding other subsequent partners will be generated using episodic memory. For more details on our adopted methodology, please refer to the Transparent methods section of the Supplemental information.

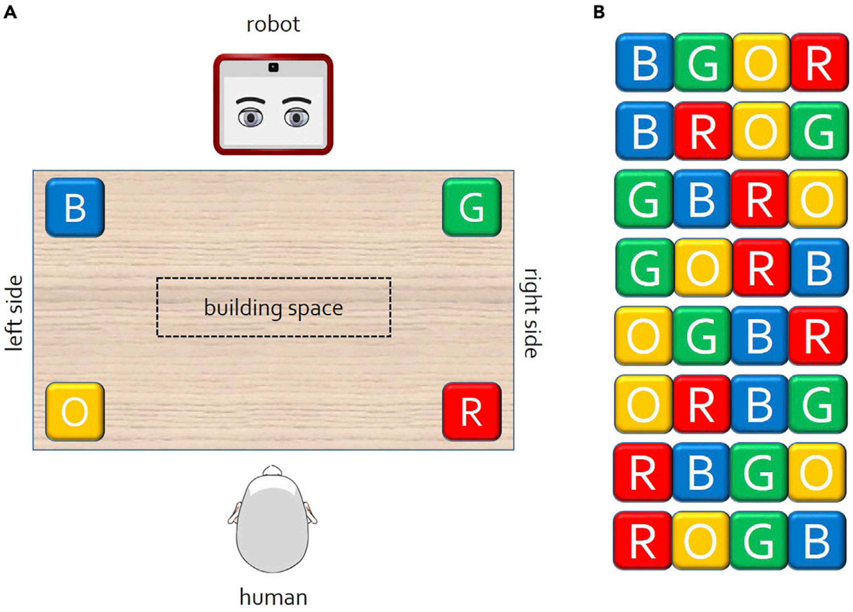


Figure 3. Experimental setup for the block building game

(A) Schematic of the playing table, depicting the position of the 4 colored blocks: blue (B), orange (O), red (R), and green (G).

(B) The 8 admissible block sequences obtained by picking blocks alternatively from each side. These sequences are the goals for this scenario.

During the execution phase, the robot will follow the workflow described in the Transparent methods (Section S1.3). In our setting, a total output represents a full line of 4 colored blocks, while the partial output (PO) is the sequence of cubes that the human has arranged before the artificial agent was able to perform intention reading. If the human is trusted or the PO is valid, the robot will collect the next predicted blocks and hand them over to him or her. If not, the robot will position the blocks itself on the building area in what it considers to be the correct order, attempting to rectify the errors that have been committed. In the latter case, the robot will also offer an explanation of why it thinks the PO is invalid (in our experimental setting, this happens when two blocks from the same side of the table are placed one next to another).

In the scope of this experiment, an interaction will be considered successful if its outcome is a structure that follows the game's rules, in other words one of those listed in Figure 3B. This is true even if the true goal was not the one predicted by the robot: this is because we do not wish to measure the performance of the intention reading model (which has already been quantified) but rather we want to evaluate the collaborative effort itself. From here on, we will define a "positive" interaction one in which the human correctly achieves a valid goal and a "negative" one where he or she takes an unsuccessful course of action. The human might violate the rules more or less intentionally, but for our purposes, we consider both these cases as a failure that will lead to a decrease of their trust level.

To verify and measure the trust model's impact on the collaborative effort driven by the intention reading architecture, we have conducted a batch of simulated experiments (The use of virtual agents in a simulated environment is a COVID-19 lockdown contingency choice) using a virtual robot which has been modeled in accordance to the empirical data collected during our latest experiment on intention reading (Vinanzi et al., 2020).

After training the robot, we let it interact with a set of simulated humans which possess different behavior patterns. It is important to note that in most of these experiments we do not make an explicit use of episodic memory. This is because, having only familiarized with the demonstrator, the robot would generate a fully trustful network for the novel informant because it will be sampling episodes from a batch of positive memories. This mean that, for the purpose of the simulated experiment, we can simply assume that the robot will naively trust its new informant. Thereafter, we continue not using the memory system because we do not want our results to depend on the order in which the robot has experienced the users, rather we want to study how each robot would respond to each user independently. For completeness, one of our simulated humans is initialized with a distrustful BN to simulate the effects of the episodic memory.

We have divided the simulated humans in two groups. The first one involves the ''deterministic'' agents, which have a fixed behavioral pattern, as follows:

- H_1: always negative;
- H_2: 50% positive, then 50% negative;
- H_3: 50% negative, then 50% positive;

The second group categorizes the ''stochastic'' agents: the latter possess different success-to-failure ratios, but the order of their actions is randomized and not fixed. In particular, we have the following:

- H_4: 50% success rate;
- H_5: 80% success rate;
- H_6: 20% success rate;
- H_7: 80% success rate, but initialized with a distrustful BN;

The deterministic humans have been tested through a batch of 100 iterations each. For the stochastic ones, we have performed 10 random initializations, and for each of them, we have executed 100 interactions with the simulated robot. The only exception is H_4, for which we performed 20 random initializations due to its high variance. During each test, we have recorded the success rate and the opinion value, both of which are described in the following section.

Evaluation metrics

Success rate

Given a human partner H_i, we define the success rate S as follows:

$$S(H_i) = \frac{\text{successful goals}}{\text{total interactions}} \in [0, 1] \qquad \text{(Equation 1)}$$

We wish to formulate a comparison between the integrated cognitive architecture and the NTA. To do so, we refer to the success rate calculated on the latter as $S^\star(H_i)$ and we formalize the difference between the two systems as follows:

$$\Delta S(H_i) = S(H_i) - S^\star(H_i) \qquad \text{(Equation 2)}$$

Positive values of $\Delta S(H_i)$ will denote a more performative collaboration obtained by our current architecture over the NTA and vice versa.

Artificial opinion

We define a quantitative index which reflects the willingness of the robot to change its opinion about a partner. For a partner H_i at a certain time step t, this artificial opinion is calculated as follows:

$$O(H_i, t) = \frac{n_p - n_n}{n_p + n_n} \in [-1, 1] \qquad \text{(Equation 3)}$$

where n_p and n_n indicate, respectively, the number of positive and negative episodes experienced by the robot with partner H_i at time t. We will sometimes use a more simple notation, where we indicate the opinion of a robot toward a generic partner at a certain timestep simply as $O(H)$.

When the robot trusts the person, that is, when $P_{X_I}(a) > P_{X_I}(b)$, it is also true that $O(H) \geq 0$ and vice versa, when the BN is distrustful toward them $O(H) < 0$. The choice of having the robot to trust when $P_{X_I}(a) = P_{X_I}(b)$ and $O(H) = 0$ is made by design since we wish the robot to act more friendly toward its users, giving them the benefit of doubt. The closest $O(H)$ is to 0, the easier it will be for the agent to flip its trust and vice versa, and the more this value tends toward the extremes, the less inclined the robot will be to alter its

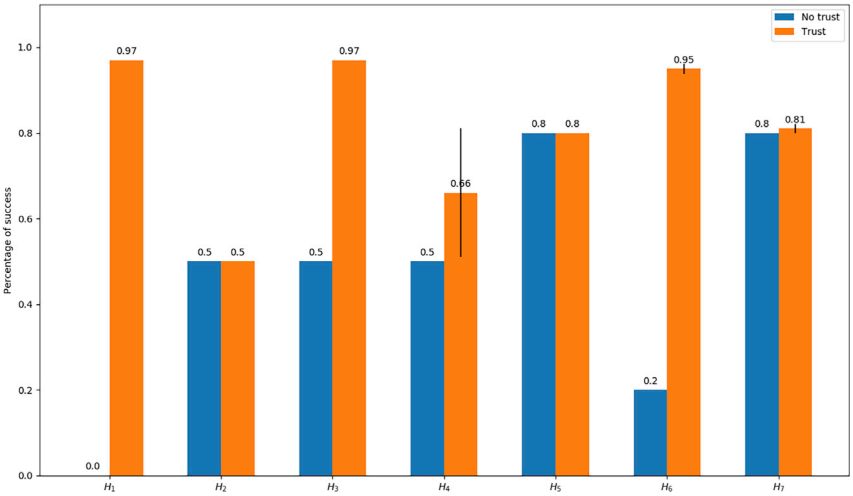


Figure 4. Comparison of the collaboration success rates with and without the trust model for each of the simulated informants

belief. Of course, $O(H) = \pm 1$ indicates a very strong opinion and it is possible only when the agent has experienced solely positive or negative episodes with that specific user.

Success rates

In our last experiment on intention reading (Vinanzi et al., 2020), we have considered partners which always act toward one of the correct goals. This means that for a hypothetical human H_0 acting always positively, $S(H_0) = S^{\bigstar}(H_0) = 1.0$ despite the fact that the empirical results we collected during that experiment indicate that the robot succeeds 80% of the time: this is because in the current investigation, we are not testing the intention reading capabilities, which enable the collaboration in the first place, rather we want to analyze the effect of a trust mechanism to correct partners who are not capable or willing to achieve a valid goal.

However, if we start considering humans which can (more or less intentionally) fail the task, the NTA's success rate drops drastically as it does not possess the ability to adopt any corrective actions. In this case, each action failed by the human will result in a failed collaboration. Figure 4 shows a comparison between the success rates of the two architectures measured on the 7 simulated humans. H_1 always fails the task so $S^{\bigstar}(H_1) = 0$, while the trust-enabled model is able to score $S(H_1) = 0.97$ with a significant increase of $\Delta S(H_1) = 0.97$.

Both H_2 and H_3 provide a mixed scenario in which the behavior of the simulated human is quite regular by being respectively positive and negative for half of the time, in inverse order. In both these cases, the NTA could only score $S^{\bigstar}(H_2) = S^{\bigstar}(H_3) = 0.5$. The trust mechanism did not prove itself of much use for H_2 since the robot builds up a strong trust for the user and is not able to change its mind in time to correct the new behavior: as we will see in the next section, this is because the agent should be observing at least $n_p + n_n + 1$ negative cases to completely change its mind about the informant, which is not possible in this 50-50 split case initialized with positive episodes. In summary, $S(H_2) = 0.5$ and $\Delta S(H_2) = 0$, in other words the performance is the same as the one obtained through NTA. H_3 behaves similarly: not having enough time to change its mind, the robot continues to distrust the human nearly until the end. The difference is that in this condition the robot maintains a strict supervision on the interactions, leading to $S(H_2) = 0.97$ with an increase of $\Delta S(H_3) = 0.47$.

To better evaluate the stochastic humans, we have recorded the success rates achieved through the batches of random initializations and we have calculated the mean μ and the standard deviation σ. The success rates reported in Figure 4 for these simulated people represent the mean score, supplied with error bars representing σ. These values are also recorded in Table 1 for better visualization.

H_4 is the agent who achieved the highest σ, that is because its behavior is the most unpredictable. This is explainable by considering what this behavioral pattern represents: with 50 positive and 50 negative

Table 1. Mean and standard deviation of the success rates calculated on the interactions performed by the stochastic simulated humans

Partner	Mean (μ)	Standard deviation (σ)	Initializations
H_4	0.66	0.15	20
H_5	0.8	0.0	10
H_6	0.95	0.01	10
H_7	0.81	0.01	10

episodes with randomized order of appearances, the trust levels can fluctuate significantly. This is also the reason behind our decision to execute double the number of trials with this simulated human. In this case, the NTA would have achieved $S^{\star}(H_5) = 0.5$, but the trust-enabled architecture is able to score $S(H_2) = 0.66$, with $\sigma = 0.15$, achieving on average $\Delta S(H_4) = 0.16$. The performance of the TA has a theoretical lower bound equal to the one obtained by the NTA and in fact we have registered scores per batch not lower than 0.5, up to a maximum of 0.93. We can conclude that a success rate of 50% is a critical point of uncertainty in which the human's behavior is too variable for the robot to adapt efficiently. As we will see shortly, above this value, the human becomes more skilled and the value of trust-based corrective mechanisms gradually fades away and vice versa, lower success rates benefit more from the TA.

H_5 is a fairly expert human who succeeds 80% of the time, which means that $S^{\star}(H_5) = 0.5$. The robot builds a very solid trust toward this partner, at the point that the 20 failures are, in our experiments, sufficiently sparse in the set of 100 interactions to never make the trust flip to negative. The latter is of course theoretically possible, but they should appear clustered at the beginning of the batch to make that occur. This means that the robot never looses trust toward this confident human but that also those 20% failures are not being captured, hence $S(H_5) = 0.8$, $\Delta S(H_5) = 0$, and $\sigma = 0$. This result is quite important because, as we mentioned previously, it demonstrates that the overall effectiveness of trust evaluations on the collaboration is inversely proportional to the skill of the partner.

The behavior of H_6 is quite the opposite of H_5, succeeding only 20% of the times. In this case, $S^{\star}(H_6) = 0.2$, but the full architecture was quickly able to detect the negative attitude of this simulated human and it promptly started distrusting them, achieving $S(H_6) = 0.95$ with $\sigma = 0.01$, leading to an average $\Delta S(H_6) = 75$.

H_7 has the same behavioral pattern than H_5, which is an 80% success rate, but the robot facing him or her is not initialized with a trusting BN but rather with a naively distrustful network. This is meant to test the effects of the episodic memory on the performance of the architecture. As we can see from Figure 4, we achieve a similar result as H_5, just slightly better because the robot will tend to not trust them and take over the task until it is persuaded about their skill. The mean result for this scenario is $S(H_7) = 0.81$, with $\sigma = 0.01$ and $\Delta S(H_6) = 0.1$. What this result stands for is the fact that the episodic memory has only a local effect on the robot's behavior, which is tuned on the long term through real interactions which take over its initial prejudice.

Overall, the experiments showed an average success rate increase equal to the following:

$$\frac{1}{7}\sum_{i=1}^{7} \Delta S(H_i) = 0.33 \qquad \text{(Equation 4)}$$

Thus confirming the positive impact of trust estimation in support of intention reading during collaborative HRI.

Trust dynamics

During the simulated interactions, we have recorded the opinion value for each of the human partners. By a design choice, the robot is initialized with a trustful BN built from 4 positive episodes. This network yields an initial opinion $O(H,0) = 0.4$. After that, we recorded $O(H,t)$ for $t \in \{1, 100\}$ and we reported them in a set of graphs.

Figure 5 shows the dynamics of the robot's opinion through the various interactions for the deterministic humans. H_1 always acts incorrectly, but the network is initially willing to trust them. This changes very quickly

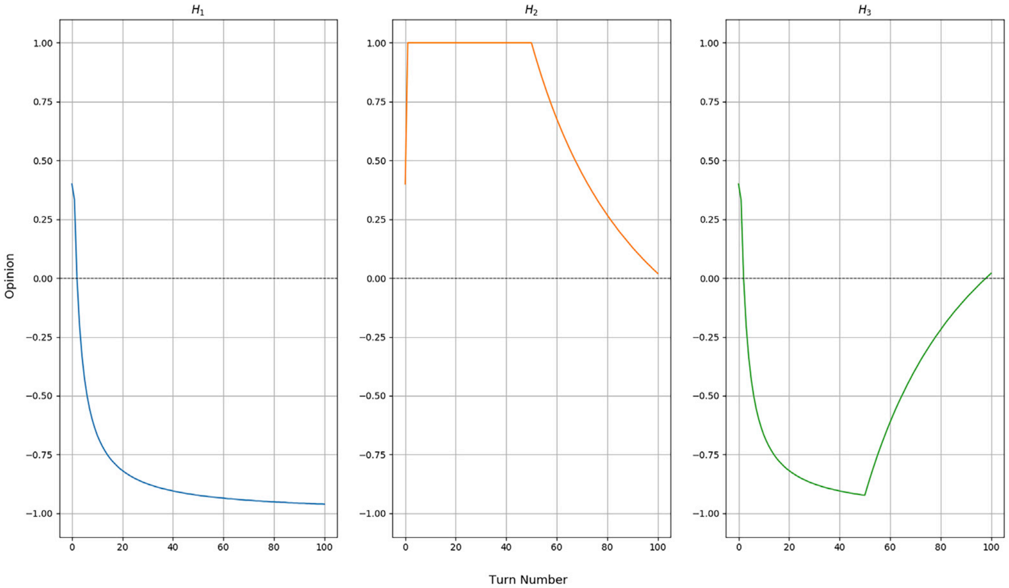


Figure 5. Variation of the opinion value at each turn of interaction for the 3 deterministic simulated informants (H_1, H_2, and H_3), initialized with a trusting BN

When $O(H)$ becomes less than 0, the robot starts distrusting the informant and taking more control on the task.

since we can observe the opinion dropping to 0 after only a few negative episodes and then decreasing close to the lower limits. This value never actually reaches the minimum value of -1 because this would only be possible if the robot had experienced 100% negative episodes, which is not the case due to how its BN was initialized. In any case, we can see how the opinion of the robot stays low, meaning that the human will have to put a lot of effort to regain its trust.

H_2 behaves 50 times positively and, subsequently, 50 times negatively. During the first half of the interactions, the opinion raises to its maximum since the robot has only experienced successful interactions with that person. From turn 51 onward, the human starts failing the block building game and the opinion slowly decreases but it is not able to flip. This is because, by the end of the session, the robot possesses 54 positive and 50 negative episodes in its memory, meaning that it has not enough time to change its mind (other 4 negative episodes will bring the opinion to 0 and another one after that will flip the trust).

H_3 acts in the opposite way as for the previous simulated agent. The trust quickly drops in the distrusting side of the graph and slowly rises after turn 50. In contrast with H_2, this human is able to flip the trust back to positive by the end of the session because of the way it was initialized. If the BN was originally set to distrust, these two graphs would result inverted.

As previously mentioned, the random nature of the stochastic humans required several batches of iterations, performed with different random initializations, to fully understand the behavior of each agent.

Figure 6 reports the dynamics of the robot's opinion during 10 out of the 20 iterations performed for H_4, which is the simulated agent with a success rate of 50%. What is immediately noticeable from these graphs is that the opinion always converges around 0: this is an expected result since this value is the midpoint in the scale, representing partners with mixed, indecisive behaviors. It is worth remembering that the robot will trust a human when $O(H) \geq 0$.

Regarding H_5, having a success rate of 80%, we expected the robot to terminate each iteration with a high opinion. This prevision was confirmed by the graphs reported in Figure 7, which show that the robot never fully changed its impression of the partner, in other words the 20 errors randomly scattered among the 100 interactions were not sufficient to flip the trust. The closest the robot got to distrust them happened in the

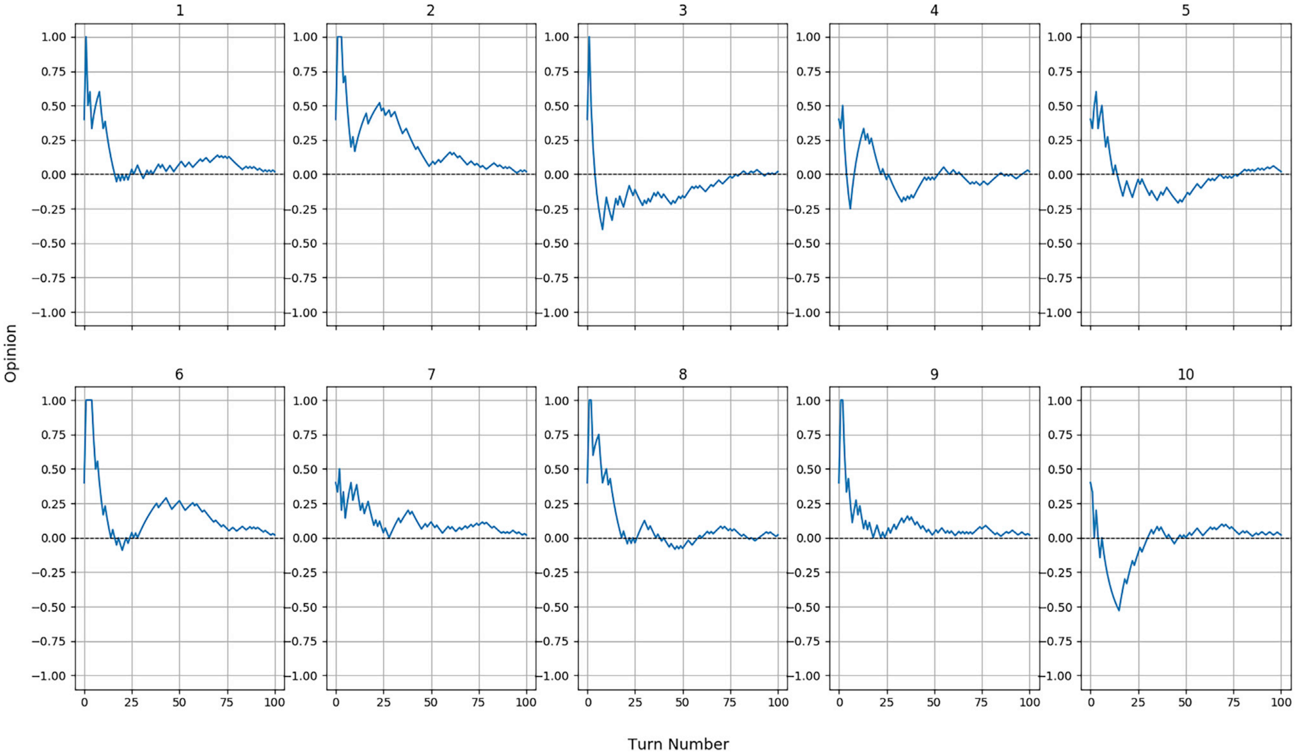


Figure 6. Opinion dynamics for the stochastic human H_4 (50% success rate) during its first 10 iterations

first diagram, where a sequence of negative episodes were experienced right at the beginning, dropping the opinion to 0, which by our design still represents a trusting situation.

Similar considerations are valid for H_6, the virtual agent capable of a very low, 20% success rate. The 10 diagrams of Figure 8 differ mostly on the very first interactions, when a sequence of positive episodes may impact the limited memory of experiences of the robot and in fact some of the iterations have managed to achieve trust for some turns. Ultimately, the opinion always ends up settling on the lower side of the graph, in the distrust domain, which is what we expect from a human who consistently fails the majority of the tasks.

All the previous simulations have been executed on a simulated robot initialized with a trustful BN, for the reasons we have explained in the preceding sections. We now wish to analyze what would happen if the network was created through episodic memory, that is, if it does not contain 4 positive episodes but a certain number of negative ones. For this reason, we have built H_7 with a BN composed of 4 negative episodes: this yields $O(H_7,0) = -0.4$. Figure 9 shows the result of this experiment, which is comparable to the one performed for H_5 since these two simulated humans behave in the same way, with the only difference being the initial prejudice. Despite the variance in the early interactions, which can make the opinion oscillate quite widely, on the long run, the latter settles for similar values registered for H_5. This demonstrates that the episodic memory can create a local effect which influences strongly the early interactions of the robot with a person but that fades gradually once the actual experience takes over the initial prejudice. This is exactly how the episodic memory system was intended to operate. Having tested the two types of BN that can be generated by the episodic memory system (completely polarized toward trust or distrust), we do not feel the need to investigate the cases which lie in between: these will produce similar, but more mitigated, effects than the ones we have observed.

DISCUSSION

Collaboration between people has been, through history, the key to obtain the grand achievements of the human species. In a future world where humans and robots will be living closely, we want to be able to

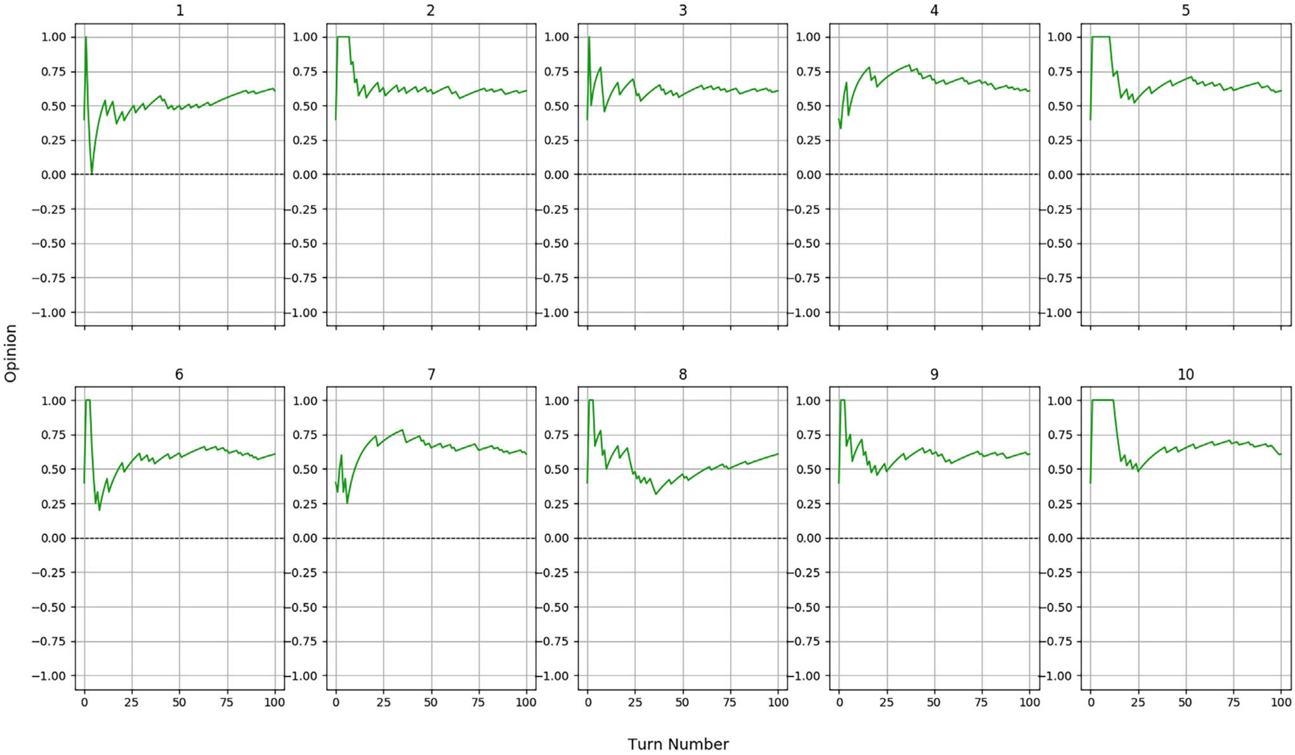


Figure 7. Opinion dynamics for the stochastic human H_5 (80% success rate) during 10 iterations

collaborate with them too. With this purpose in mind, we state that a true collaborative robot able to operate in human-sized environments must possess the same cognitive skills that drive our own social life. In this paper, we defined collaborative intelligence as the mutual interaction between intention reading and trust estimation, two mental abilities toward which humans are biologically oriented. The former allows an agent to understand the actions and goals of other agents acting around it, thus providing clues and meaning to simple sensory perceptions, while the latter is essential to estimate the level of skill or knowledge of another agent so to formulate an appropriate plan. Following the developmental robotics principles, our cognitive architecture takes inspiration from scientific findings in human cognition, and both the intention reading and trust models are designed according to the current psychological literature. We have developed a cognitive system which is able to learn goals by demonstration in an unsupervised and probabilistic way and to estimate trust using an artificial ToM. We have applied this architecture to a block building game where a robot is engaged with several humans to pick and place some colored cubes from a table to form constructions that obey to certain patterns.

Overall, we can conclude that the synergistic combination of intention reading and trust leads to better results than the ones obtainable by just predicting the human's goal. The experiments that we conducted have shown that the complementary use of both these cognitive skills enhances the collaborative performance, making the robot act as a better teammate. This confirms our initial hypothesis, which is that collaborative intelligence is enabled by the ability to read another agent's intention and is fostered by the capacity to correctly estimate the trustworthiness of the other party. The robot's ability to take control of the task whenever the partner demonstrates a lack of skill results in a significant increase in the success of the joint task.

Both the intention reading and the trust models offer directions in which to orient future investigations. The former, for example, could benefit from the addition of hierarchical goals, i.e., goals composed by multiple sub-goals (for example, uncapping a bottle might be one step to achieve the "drink" goal). Another possible study could explore the use of more social clues and the investigation of their order of application within Feature-Space Split Clustering, the multi-modal clustering algorithm which we use within the

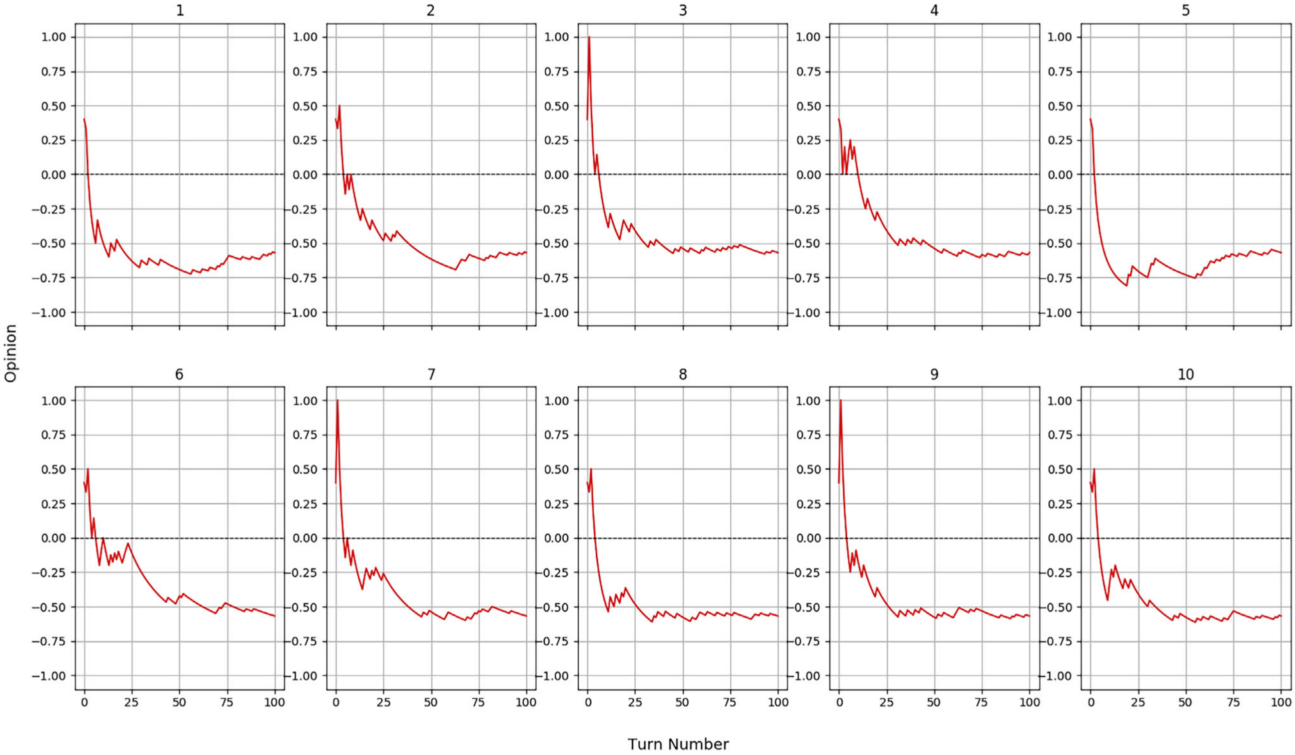


Figure 8. Opinion dynamics for the stochastic human H_6 (20% success rate) during 10 iterations

intention reading module, described in detail in the Supplemental information. Both these components will be revised in the near future to apply them to multi-agent systems: ensembles of heterogeneous agents, each of which has the ability to contribute to a greater problem-solving network. In this kind of scenario, it would be possible to take into account the contemporary influence of two or more agents, similarly to what has been done by Butterfield et al. (2009). This, of course, would also imply the adaptation of this cognitive architecture to collaborate not only with humans but also with other artificial agents.

Another future direction could involve the use of this architecture within a more continuous representation of trust, where a partner possesses a degree of trustworthiness as opposed to a binary state. Having access to a more refined representation could provide further benefits for the robot: for example, this could translate in a continuous definition of the collaboration process, where the agent might decide to take over only a subset of the actions based on their complexity.

LIMITATIONS OF THE STUDY

Due to the inability to access appropriate research facilities due to COVID-19 lockdown in the United Kingdom, it was not possible to perform experiments on the physical robot, which were instead replaced by simulations. By providing the virtual robot the same empirical error rates obtained during the foundational experiments, which have all been executed in the real world, we have tried to minimize any approximation errors between the simulated and the real interactions.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Samuele Vinanzi (samuele.vinanzi@manchester.ac.uk).

Material availability

This study did not generate new unique materials.

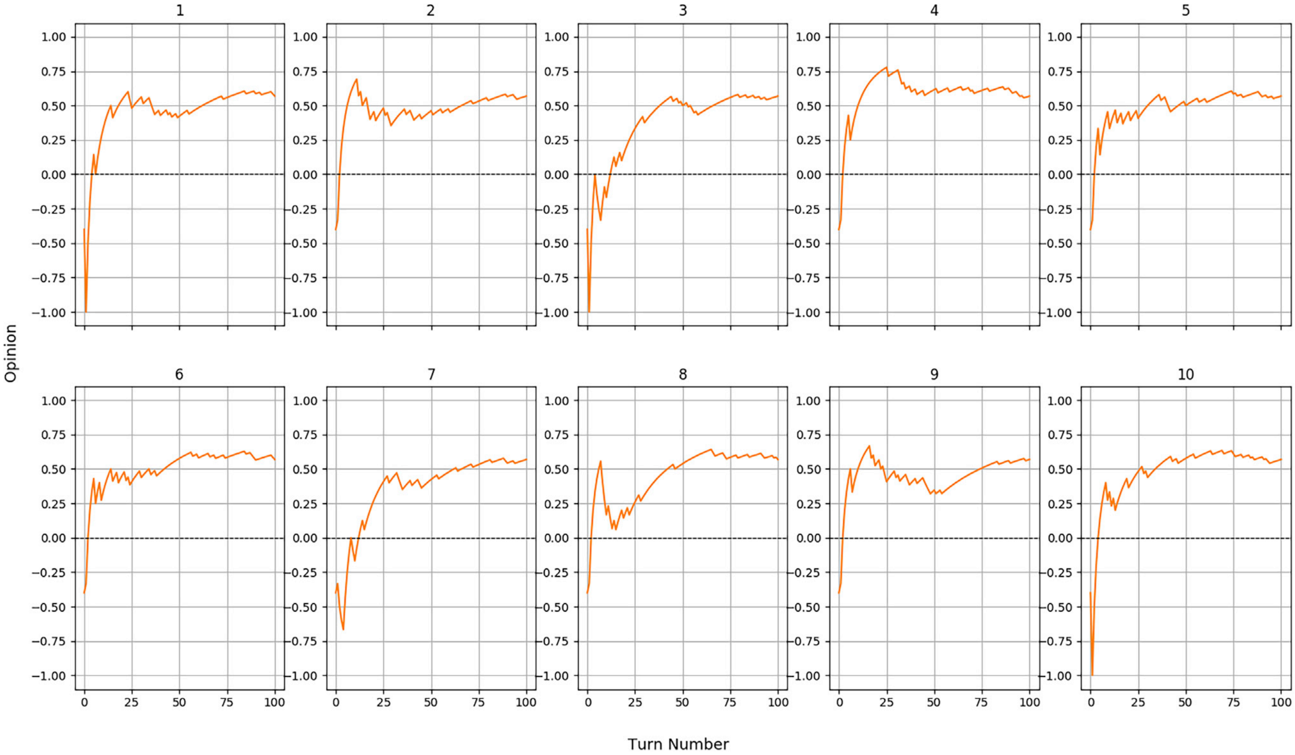


Figure 9. Opinion dynamics for the stochastic human H_7 (80% success rate, against a naively distrusting BN) during 10 iterations

Data and code availability

The data and code in this manuscript are available at github.com/samvinanzi/DeCIFER.

METHODS

All methods can be found in the accompanying Transparent methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2021.102130.

ACKNOWLEDGMENTS

This work was supported by a project funded by the Honda Research Institute Europe (DeCIFER project) and the material is in part based upon work supported by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF under Awards FA9550-19-1-7002.

AUTHOR CONTRIBUTIONS

S.V. designed the architecture, implemented the code, ran the experiments, and wrote the manuscript. C.G. and A.C. supervised the research and provided insights and guidance.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 16, 2020
Revised: January 7, 2021
Accepted: January 26, 2021
Published: February 19, 2021

REFERENCES

Baldwin, D.A., and Baird, J.A. (2001). Discerning intentions in dynamic human action. Trends Cogn. Sci. *5*, 171–178.

Bauer, A., Wollherr, D., and Buss, M. (2008). Human–robot collaboration: a survey. Int. J. HR *5*, 47–66.

Butterfield, J., Jenkins, O.C., Sobel, D.M., and Schwertfeger, J. (2009). Modeling aspects of theory of mind with markov random fields. Int. J. Soc. Robot. *1*, 41–51.

Das, T., and Teng, B.-S. (2004). The risk-based view of trust: a conceptual framework. J. Bus. Psychol. *19*, 85–116.

De Castro, E.C., and Gudwin, R.R. (2010). An episodic memory for a simulated autonomous robot. Proc. Robocontrol *2010*, 1–7.

Dillenbourg, P. (1999). What do you mean by collaborative learning? In Collaborative-learning: Cognitive and Computational Approaches (Elsevier), pp. 1–19.

Dindo, H., Donnarumma, F., Chersi, F., and Pezzulo, G. (2015). The intentional stance as structure learning: a computational perspective on mindreading. Biol. Cybern. *109*, 453–467.

Dominey, P.F., and Warneken, F. (2011). The basis of shared intentions in human and robot cognition. New Ideas Psychol. *29*, 260–274.

Ebstein, R.P., Israel, S., Chew, S.H., Zhong, S., and Knafo, A. (2010). Genetics of human social behavior. Neuron *65*, 831–844.

Erikson, E.H. (1993). Childhood and Society (W. W. Norton & Company).

Ferrari, P.F., Tramacere, A., Simpson, E.A., and Iriki, A. (2013). Mirror neurons through the lens of epigenetics. Trends Cogn. Sci. *17*, 450–457.

Floyd, M. W., Drinkwater, M., and Aha, D. W. (2014). Adapting autonomous behavior using an inverse trust estimation. In International Conference on Computational Science and Its Applications, pages 728–742. Springer.

Gallese, V., and Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. Trends Cogn. Sci. *2*, 493–501.

Gill, Z. (2012). User-driven collaborative intelligence: social networks as crowdsourcing ecosystems. In CHI'12 Extended Abstracts on Human Factors in Computing Systems, pages 161–170.

Granada, R., Pereira, R. F., Monteiro, J., Barros, R., Ruiz, D., and Meneguzzi, F. (1995). Hybrid activity and plan recognition for video streams. In The 2017 AAAI Workshop on Plan, Activity, and Intent Recognition. Greer, JE and Koehn, GM, pages 54–59.

Groom, V., and Nass, C. (2007). Can robots be teammates?: Benchmarks in human–robot teams. Interact. Stud. *8*, 483–500.

Henderson, A.M., and Woodward, A.L. (2011). "let's work together": what do infants understand about collaborative goals? Cognition *121*, 12–21.

Jansen, B., and Belpaeme, T. (2006). A computational model of intention reading in imitation. Robot. Autonomous Syst. *54*, 394–402.

Jockel, S., Weser, M., Westhoff, D., and Zhang, J. (2008). Towards an episodic memory for cognitive robots. In Proc. of 6th Cognitive Robotics workshop at 18th European Conf. on Artificial Intelligence (ECAI), pages 68–74.

Jones, G.R., and George, J.M. (1998). The experience and evolution of trust: Implications for cooperation and teamwork. Acad. Manag. Rev. *23*, 531–546.

Kelley, R., King, C., Tavakkoli, A., Nicolescu, M., Nicolescu, M., and Bebis, G. (2008). An architecture for understanding intent using a novel hidden markov formulation. Int. J. HR *5*, 203–224.

Malle, B.F., Moses, L.J., and Baldwin, D.A. (2001). Intentions and Intentionality: Foundations of Social Cognition (MIT press).

Manzi, A., Dario, P., and Cavallo, F. (2017). A human activity recognition system based on dynamic clustering of skeleton data. Sensors *17*, 1100.

Mayer, R.C., Davis, J.H., and Schoorman, F.D. (1995). An integrative model of organizational trust. Acad. Manag. Rev. *20*, 709–734.

Patacchiola, M. and Cangelosi, A. (2016). A developmental bayesian model of trust in artificial cognitive systems. In Development and Learning and Epigenetic Robotics (ICDL-EpiRob), 2016 Joint IEEE International Conference on, pages 117–123. IEEE.

Premack, D., and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? Behav. Brain Sci. *1*, 515–526.

Rizzolatti, G., and Craighero, L. (2004). The mirror-neuron system. Annu. Rev. Neurosci. *27*, 169–192.

Roschelle, J., and Teasley, S.D. (1995). The construction of shared knowledge in collaborative problem solving. Computer Supported Collaborative Learning (Springer), pp. 69–97.

Sciutti, A., Ansuini, C., Becchio, C., and Sandini, G. (2015). Investigating the ability to read othersâ€™ intentions using humanoid robots. Front. Psychol. *6*, 1362.

Singh, B., Marks, T. K., Jones, M., Tuzel, O., and Shao, M. (2016). A multi-stream bi-directional recurrent neural network for fine-grained action detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1961–1970.

Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). Understanding and sharing intentions: the origins of cultural cognition. Behav. Brain Sci. *28*, 675–691.

Tulving, E. (1972). Episodic and semantic memory. Organ. Mem. *1*, 381–403.

Vanderbilt, K.E., Liu, D., and Heyman, G.D. (2011). The development of distrust. Child Dev. *82*, 1372–1380.

Vinanzi, S., Cangelosi, A., and Goerick, C. (2020). In 2020 29th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).

Vinanzi, S., Goerick, C., and Cangelosi, A. (2019a). Mindreading for robots: Predicting intentions via dynamical clustering of human postures. In 2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob), pages 272–277.

Vinanzi, S., Patacchiola, M., Chella, A., and Cangelosi, A. (2019b). Would a robot trust you? Developmental robotics model of trust and theory of mind. Philos. Trans. R. Soc. Lond. B Biol. Sci. *374*, 20180032.

Wellman, H.M., Cross, D., and Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. Child Development *72*, 655–684.

Woodward, A.L., Sommerville, J.A., Gerson, S., Henderson, A.M., and Buresh, J. (2009). The emergence of intention attribution in infancy. Psychol. Learn. Motiv. *51*, 187–222.

Zanatto, D. (2019). When do We Cooperate with Robots?, PhD thesis (University of Plymouth).

iScience, Volume *24*

Supplemental Information

The collaborative mind: intention reading and trust in human-robot interaction

Samuele Vinanzi, Angelo Cangelosi, and Christian Goerick

1 Transparent Methods

The aim of our research is to develop a comprehensive cognitive architecture that encompasses both intention reading and trust abilities for a humanoid robot engaged in HRI. To do so, we are going to build on the foundations of our previous models (Vinanzi et al., 2020, Vinanzi et al., 2019) and build an integration that, following the schematic presented in Figure 1, will allow a robot to act collaboratively towards a human partner. This architecture will be used in a scenario in which the robot will have to infer the goal of its partner by the observation of their social cues and subsequently perform decision-making to formulate an action plan that will try to optimize the chances of successfully achieving the intended objective.

For the design of both the artificial intention reading and trust capabilities, we made use of the developmental robotics approach. Cangelosi et al. (2015) defined this subject as “the approach to the design of behavioral and cognitive capabilities in artificial agents that takes direct inspiration from the developmental principles and mechanisms observed in the natural cognitive systems of children”. In other words, our computational models are inspired by scientific findings in human cognition.

1.1 Intention Reading

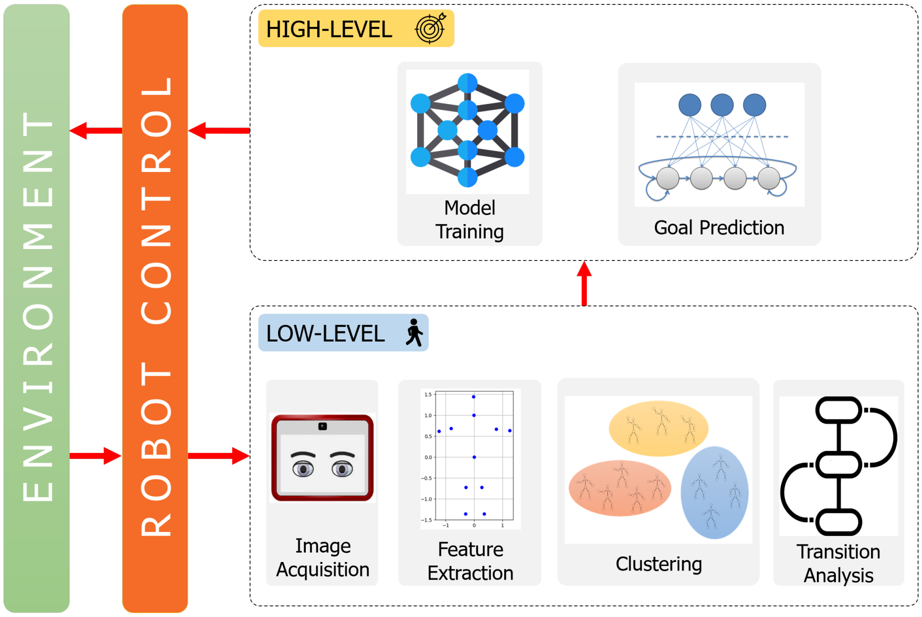
1.1.1 Motivation

In the pursuit of the developmental robotics approach, the intention reading model lacks a pre-existing plan library, rather it follows the psychological theories which state that this cognitive ability is learned by experience (Malle et al., 2001). Furthermore, it follows the principles theorized by Tomasello et al. (2005) which state that the intention decoding task is divided in a low-level action understanding based on social cues and a high-level goal prediction. The lack of a hand-crafted plan library means that the robot will be able to learn goals in a more flexible and scalable way, while the use of unsupervised an probabilistic models, rather than supervised ones such as neural networks, makes the robot learn on the fly with no need for big datasets or long traning times, making this architecture lightweight on a computational point of view.

An overview of the intention reading architecture is shown in Figure 1. Additional details about this section can be found in our previous publications (Vinanzi et al., 2019, 2020).

1.1.2 Low-level social cue clustering and action representation

The low-level module of the intention reading architecture tries to encode temporal sequences of human configurations, expressed as sets of social cues, into a more compact representation that will be used to recognize actions. The main idea is to observe the human acting to achieve the goals, then cluster the set of their social cues and analyze how their actions unravel through these clusters to form an encoding that will be used by the high-level goal prediction module.

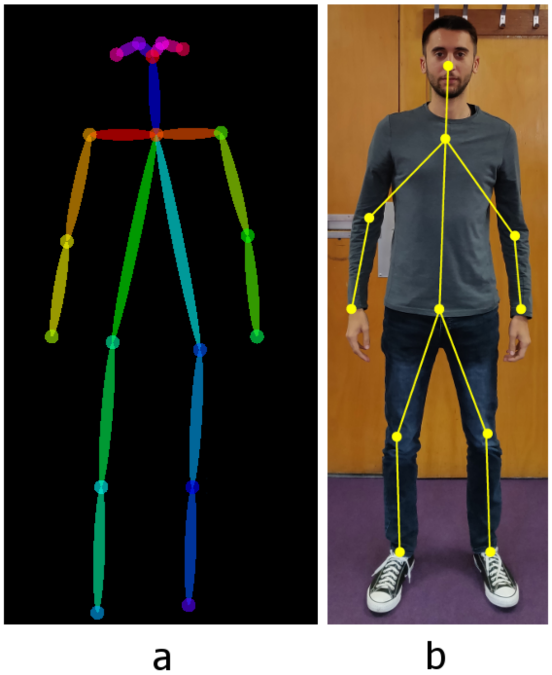


Supplemental Figure 1: Overview of the intention reading architecture. The low-level extracts social data from the optical stream, forms clusters and uses them to represent actions as transitions through clusters. The high-level uses this encoding to probabilistically infer the pursued goal. The robot control manages sensors and actuators. Related to Figure 2.

Following the psychological literature (Tomasello et al., 2005), we chose to employ body posture and eye gaze direction as the features that the robot is going to observe from the human.

We collect postural information by generating skeleton data through the use of a pre-trained deep convolutional neural network architecture named OpenPose (Cao et al., 2016), specialized in real-time multi-person 2D pose estimation. This neural network receives in input the images from the robot's eye camera and outputs a 18x2 feature vector representing the detected skeleton keypoints as 18 joints expressed in 2D spatial coordinates, as reported in Figure 2a. In order to optimize memory and speed requirements and to comply with recent findings which state that classification tasks achieve better results with a reduced set of joints (Manzi et al., 2017), we operate a keypoint reduction to diminish the volume of data required for each skeleton. To do so, we discard the keypoints corresponding to the eyes, ears and shoulders, whilst calculating a new torso keypoint as a median between the two hips: doing so, we obtain a more compact 11x2 representation shown in Figure 2b.

The skeletons generated by this procedure cannot be used directly for classification purposes, as they are dependent on the position and size of the subject.



Supplemental Figure 2: A comparison between the skeletal keypoints extracted from the camera image (a) and the reduced keypoint set computed by the system (b). Related to Figure 2.

To overcome this problem and obtain spacial invariance, we apply a normalization process introduced by Cippitelli et al. (2016). For a skeleton with n joints, the feature vector f is defined as:

$$f = [J_1, J_2, ..., J_n] \tag{1}$$

where J_i is a vector containing the normalized 2D coordinates of the ith keypoint:

$$J_i = \frac{J_i - J_0}{\|J_1 - J_0\|} \tag{2}$$

where J_0 and J_1 are, respectively, the neck and torso joint. The latter will be located on the origin of the cartesian space, so its components will all be zero. For this reason, it is removed from the feature vector, which at this point will have a dimension of 10x2: this corresponds to a 44.5% size reduction from to the original representation.

Another social cue that we collect from the robot's partner is gaze direction. We use Deepgaze (Patacchiola and Cangelosi, 2017), a convolutional neural network specialized in head pose estimation, to retrieve a 3D vector representing estimated roll, pitch and yaw of the human for each image acquired by the robot. We chose to approximate gaze direction with head orientation to avoid some computational overheads that would impair the real-time computation of several frames per second. This has been proved to be an acceptable approximation (Jha and Busso, 2017).

We create action representations from the perceptual data through an unsupervised clustering procedure, using a novel algorithm that combines multiple sets of features in several increasingly refined stages, which we call Feature-Space Split Clustering (FSSC). This strategy is adopted because it is possible to distinguish complex and potentially ambiguous actions by increasing the granularity of the clustering operation, which means taking into account a multitude of social cues. The main idea behind FSSC is a multi-level clustering process that uses only a subset of the features at each level.

Consider a set of M training samples:

$$X = \left\{ x^{(1)}, x^{(2)}, \ldots, x^{(M)} \right\} \tag{3}$$

Each sample can be seen as defined by N groups of features:

$$x^{(i)} = \left\{ f_1^{(i)}, f_2^{(i)}, \ldots, f_N^{(i)} \right\} \tag{4}$$

Each group defines the feature-space f_n with $n \in [1, N]$ and contains data extracted from a different perceptual input. In our scenario we use $N = 2$ and for each image i we have that $f_1^{(i)}$ is a 20D vector containing the skeleton keypoints configuration and $f_2^{(i)}$ is a 3D vector that specifies the gaze direction.

FSSC works by implicitly computing a tree of depth $L = N$ whose nodes contain the refined clusters. The root node ($\ell = 0$) contains all the data samples and is considered as a single cluster, while nodes of each subsequent level $\ell > 0$ are the clusters obtained by clustering the samples belonging to the parent node in the feature-space f_ℓ. At each level, we perform Principal Component Analysis (PCA) dimensionality reduction to project the data in a 2D space to avoid the curse of dimensionality (Bellman, 2013). We do so also because clustering relies on euclidean distance as a metric, but in high dimensional spaces the concept of distance becomes less precise, since it tends to converge. Finally, we chose X-Means (Pelleg et al., 2000) as the internal clustering method, which is a variation on the traditional K-Means algorithm that overcomes its principal limitation: the need to manually specify the parameter K that defines the desired number of clusters. The model selects the optimal one by performing model selection among a finite set of models through the optimization of the Bayesian Information Criterion. Algorithm 1 describes this computation.

Given the hierarchical and nonlinear nature of this algorithm, we can't perform classification (intended as the association of a new data sample to one of the existing clusters) through a simple Euclidean distance search for the closest centroid. Instead, the procedure described in Algorithm 2 must be adopted.

Algorithm 1: Feature-Space Split Clustering (FSSC)

```
Input: training samples $X$; number of feature sets $N$
Output: A tree of clusters
$tree \leftarrow \{\}$
Initialize the root node with all the samples $X$
for $\ell \leftarrow 0$ to $N$ do
    foreach cluster of level $\ell$ do
        $x \leftarrow$ samples belonging to cluster
        $f \leftarrow f^{(x)}_{\ell+1}$
        $f' \leftarrow$ Dimensionality reduction on $f$
        $newClusters \leftarrow CLUSTERING(f')$
        Set $newClusters$ to level $\ell + 1$
        $tree \leftarrow tree \cup newClusters$
    end
end
return tree
```

The latter searches through the cluster tree, comparing the centroids in their respective feature-space coordinates until a leaf node is found.

It is important to note that, despite this approach splits the feature-spaces, in fact it does not decouple the multi-modal features extracted from the human, as they share a temporal dependence on the frame from which they were generated. In other words, the features never lose their alignment.

During its initial training, the agent will observe the actions of its human partner, record the image frames and extract all the relevant features. This assembled dataset will be used to create the clusters using the method described in Algorithm 1: each of them represents a group of similar but not identical postures. These will be used in the next stage, Transition Analysis, to create the low-level encoding: each action will be represented by the sequence of the cluster ids encountered during its performance. To obtain temporal invariance, we include in the encoding only transitions through different clusters, discarding the persistence in the same group: in this way, the representation of an action won't depend on the speed of its execution. The results of this analysis, plus a set of unique names for each goal which are generated automatically by the robot, are forwarded to the high-level module to train it as described in Section 1.1.3.

After being trained, the system will be able to perform intention reading: the agent will observe its human partner during the execution of the action, each of their physical configurations will be classified to one of the known clusters using the procedure described in Algorithm 2 and the discovered id will be forwarded to the high-level module for probabilistic inference.

The low-level intention reading process is sensitive to noise because it is learning each intention and its sequence of actions from a single training ex-

Algorithm 2: Cluster classification

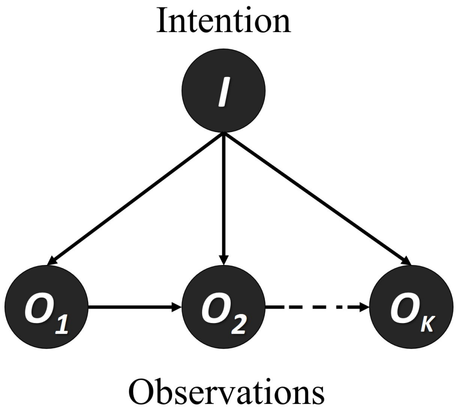
```
Input: cluster tree $T$, testing sample $s$
Output: cluster to which $s$ belongs
$parentNode \leftarrow$ root node of $T$
$\ell \leftarrow 1$
Loop
    $C \leftarrow$ descendants of $parentNode$ in $T$
    $cluster \leftarrow \min distance(f_\ell^{(s)}, c \in C)$
    if $cluster$ has descendants then
        $parentNode \leftarrow cluster$
        $\ell \leftarrow \ell + 1$
    else
        return $cluster$
    end
EndLoop
```

ample, which is in turn obtained from a non-deterministic and unsupervised process. To reduce this effect, we implement a post-processing computation that aims to ground this general architecture to our specific experiment by capturing the regularities of the data and assigning to each end position of our actions (i.e. the grasping position for each of the blocks) one of the cluster ids based on its statistical mode in the computed training dataset. Training actions are then eventually corrected by merging the two representations to avoid any errors in the training set.

1.1.3 High-level goal prediction

The high-level module is in charge of goal probabilistic inference from the observed actions. What we are trying to achieve is not action recognition but rather prediction, this means that only one part of the action will be known and observable. Our objective is to determine the intention based on as few observations as possible, so that the robot will be able to contribute to the task before it is over.

To achieve this, we have employed the BN shown in Figure 3. The top node denoted as I represents the intention of the human partner and its probability distribution is equal across all the possible goals. The bottom nodes marked as O_k with $k \in [1, K]$, where K is the maximum length of the encoded actions, represent the observations. The values of these nodes span in the range of the possible cluster ids identified by the low-level module. The conditional probability tables of the observation nodes are fitted from the training data provided by the low-level Transition Analysis (i.e. the action encoding associated to each goal name) using Maximum Likelihood Estimation (MLE) (Aldrich et al., 1997). We assume that the probability of each observation depends on the driving intention and by the precedent symbol encountered:



Supplemental Figure 3: The BN used for high-level probabilistic goal prediction. The top node represents the intention of the observed partner, whilst the bottom node symbolize the observations (the action encoding symbols produced by the low-level module). Related to Figure 2.

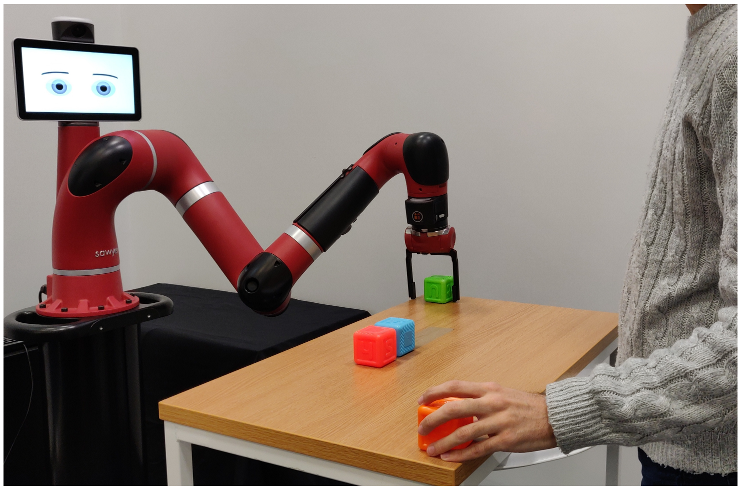
$$P(O_1 \mid I) \tag{5}$$

$$P(O_k \mid O_{k-1}, I)\colon k \in [2, K] \tag{6}$$

Once the probabilistic model is trained, it can be used for inference. During the execution phase, the robot will be observing the human and recording each cluster transition in real-time. The low-level module will forward these symbols to the high-level, which will treat them as sequential observations. Each time a new piece of evidence is added to the model, we use Pearl's Message-Passing algorithm (Lauritzen and Spiegelhalter, 1988) to calculate the marginal probability distribution for node I given the evidence. As soon as one of the goals is predicted with a probability greater than 0.5, it is sent forward in the processing chain to instantiate appropriate collaborative behavior. The value of this threshold was chosen in compliance to the time restrictions: we could choose to wait for a higher confidence, but this may slow the prediction time up to the point in which the robot's intervention in the joint task would become irrelevant.

1.1.4 Robot control

The robot control module deals with the direct interface between the cognitive architecture and the robotic platform, in this case a Sawyer: an industrial collaborative robot designed for object manipulation, equipped with a 7-DOF arm (Figure 4). In particular, it provides interaction with the ROS middleware to



Supplemental Figure 4: The Sawyer robot which was simulated for the collaborative intelligence experiments involving an interactive block placing game. Related to Figure 2 and Figure 3.

control its sensors and actuators and perform vision, movement and grasping for the shared goal task.

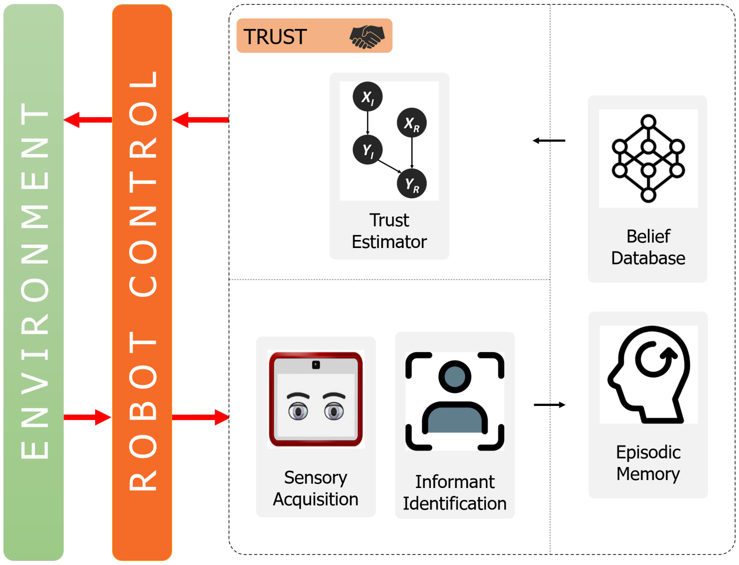
1.2 Trust Estimation

1.2.1 Motivation

The trust model has been designed to be able to reproduce the psychology experiment on ToM maturity by Vanderbilt et al. (2011). In the latter, 90 preschool-age children equally divided in 3-, 4-, and 5-years-olds were exposed to a video in which an adult actor gave advice to another adult who was trying to locate a sticker hidden in one or two boxes. The informant could be either a helper or a tricker, suggesting respectively the correct or the wrong location. In the second phase of the experiment, a child would be involved in the game and would receive the same kind of suggestion by the informant. Based on the children's choices and on some meta-cognitive questions submitted to them, Vanderbilt theorized that only the 5-year-olds were able to differentiate the helpers from the trickers, therefore demonstrating to possess a mature ToM.

1.2.2 Bayesian approach in trust reasoning

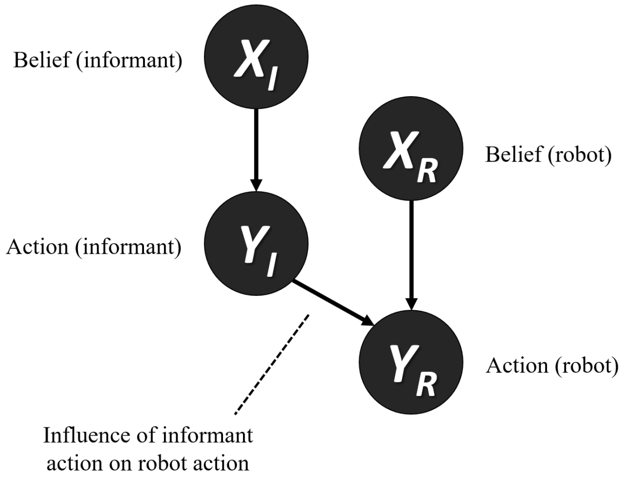
In our previous research (Vinanzi et al., 2019) we developed a probabilistic model that could allow a robot to take part in the same sticker finding experiment and to act as a child with mature or immature ToM, learning to predict the beliefs and attitudes of the informants. The trust estimation architecture



Supplemental Figure 5: Overview of the trust estimation architecture. The robot uses its sensors to identify each human and select their own BN or eventually generate a new one using episodic memory. The selected model is then used for inference. Related to Figure 2.

is reported in Figure 5. We employed a BN using discrete Boolean variables that assume two states: a and b, each corresponding to one of the two positions where the stickers can be located in the experiment. A graphical illustration of this BN can be observed in Figure 6: the two nodes X_R and Y_R represent respectively the beliefs and actions of the robot. The posterior distribution of the node Y_R allows the agent to choose the action to perform: that means searching for the sticker in position a or b. The connection between Y_I and Y_R represents the influence that the opinions of the informant have on the agent's action. The action of the agent is then a consequence of its own belief X_R and the informant action Y_I. Lastly, the estimation of X_I, the informant's belief, makes the agent able to effectively discriminate a trickery from a non-malevolent human error. The cognitive architecture we designed creates one of these BNs for each human it interacts with and uses it to predict their future behavior. Every partner is detected and recognized using Haar Cascade (Viola and Jones, 2001) and Local Binary Pattern Histogram (Ojala et al., 2002) on the robot's camera live stream.

For our current purposes, we intend to employ this model to check whether the human has the knowledge or skill to achieve a given goal: if this is not the case, the robot will have to perform corrective actions to ensure the success of the task. To do so, we have employed the same Bayesian network changing the meaning of its binary nodes: a will represent a correct goal whilst b will symbolize an incorrect goal. Following this convention, X_I and X_R will



Supplemental Figure 6: The BN that models the relation between the robot and an informant. The agent generates a separate network for each user, with the same structure but different probability distribution. Related to Figure 2.

represent, respectively, knowledge of the informant and of the robot about the correct goals, Y_I will symbolize the choice of a correct or incorrect action by the informant and finally Y_R will depict whether the robot should adopt a trusting or a corrective action.

The trust model is built from episodes, which are data structures that encode interaction outcomes. Once the agent has collected a certain amount of episodes from an informant, it can generate a BN associated to him or her using MLE to determine the conditional probability tables of its nodes. For the root nodes X_I and X_R we calculate these probabilities as:

$$\begin{aligned} P_Y(a) &= \theta \\ P_Y(b) &= 1 - \theta \end{aligned} \tag{7}$$

Denoting N_a and N_b as the number of times the human demonstrates a or b, we can estimate θ as:

$$\hat{\theta} = \frac{N_a}{N_a + N_b} \tag{8}$$

For the nodes Y_I and Y_R, instead, we have to also take into consideration the influence of the parents.

Once a BN has been created for a certain user and its parameters have been learned from the interactions, it is possible to infer the posterior probability of the nodes given some observations. We calculate posterior distributions using Pearl's Message-Passing algorithm (Lauritzen and Spiegelhalter, 1988).

The outcome of each interaction is saved as a new episode that will modify the parameters of the probabilistic model. This means that while experiencing new interactions, the BN can acquire new statistical data and adapt its behavior over time, eventually switching between trust and distrust.

1.2.3 Episodic Memory

The power to use one's own past memories to take decisions in the present and future is an important ability that enhances the cognitive processes. In the original experimental design by Vanderbilt et al. (2011), the child (or, in our case, the robot) would familiarize with the partner before the real interaction. We make use of episodic memory to let the artificial agent be able to instantly interact even with unfamiliar people. On a technical level, the main problem is to generate on the fly a new BN with adequate parameters to use with that unknown person. These parameters will depend on the robot's personal character which, in turn, depends on the way it has been treated in the past: an agent which has often experience human failures would learn to be mistrustful and vice versa, as in the "trust vs mistrust" phase in child development (Erikson, 1993).

The design guidelines that we followed in the creation of our algorithm were the following: memories fade away with time, the details become blurred proportionally to the amount of memories possessed and, finally, shocking events such as surprises and betrayals should be more difficult to forget. Our algorithm draws inspiration from the particle filter technique widely used in mobile robot localization (Rekleitis, 2004). Whenever an unknown informant is met, this component generates on the fly a certain number of episodes to train a new BN.

We define the set of BNs memorized by the agent as:

$$S = [s_0, s_1, ..., s_n] \tag{9}$$

Where n is the number of humans known by the agent.

Each BN s_i was generated by a set of episodes, and these are going to be denoted as *replay datasets* for that BN:

$$E_{s_i} = [\varepsilon_0^{(s_i)}, \varepsilon_1^{(s_i)}, ..., \varepsilon_m^{(s_i)}] : s_i \in S \tag{10}$$

Where m is equal to the number of episodes of the replay dataset. So, in this notation $\varepsilon_j^{(s_i)}$ represents the j-th episode of the replay dataset that formed the BN s_i.

The equation we are about to introduce uses information theory to quantify the amount of information each specific episode represents. Our goal is to find how much this value differs from the total entropy of its replay dataset: a high difference means that the event is to be considered surprising and must be easier to recall than ordinary, unsurprising events. For example, if a person who is always been trustful suddenly tricks the agent, this betrayal will be remembered with a greater impact. At the same time, all of the memories are subject to a

progressive time degradation that tends to blur them with a timing dependent on their importance.
Formally, a real factor denoted as importance value v defined in the interval $[0, 1]$ is calculated for every episode $\varepsilon_j^{(s_i)}$ as the difference between the amount of information of the episode, $I(\varepsilon_j^{(s_i)})$, and the total entropy of its replay dataset, $H(E_{s_i})$, divided by the discrete temporal difference from the time when the memory was formed.

$$\begin{aligned} v(\varepsilon_j^{(s_i)}) &= \frac{\mid I(\varepsilon_j^{(s_i)}) - H(E_{s_i}) \mid}{\Delta t + 1} \\ &= \frac{\mid -\log_2 P(\varepsilon_j^{(s_i)}) + \sum_{\varepsilon \in E_{s_i}} P(\varepsilon) \log_2 P(\varepsilon) \mid}{t_{present} - t_{\varepsilon_j^{(s_i)}} + 1} \end{aligned} \tag{11}$$

Equation 11 is used to weight every episode from each replay dataset in the agent's memory in order to perform a systematic resampling (Douc and Cappé, 2005) to pick the new episodes that will form the replay dataset for the new BN we intend to create, $E_{s_{n+1}}$. Finally, MLE is applied to the new reply dataset to evaluate the parameters of the network. This new BN will be stored in the agent's long term memory as s_{n+1} and will be used to predict the trustworthiness of the new informant.

1.3 Integrated Cognitive Architecture

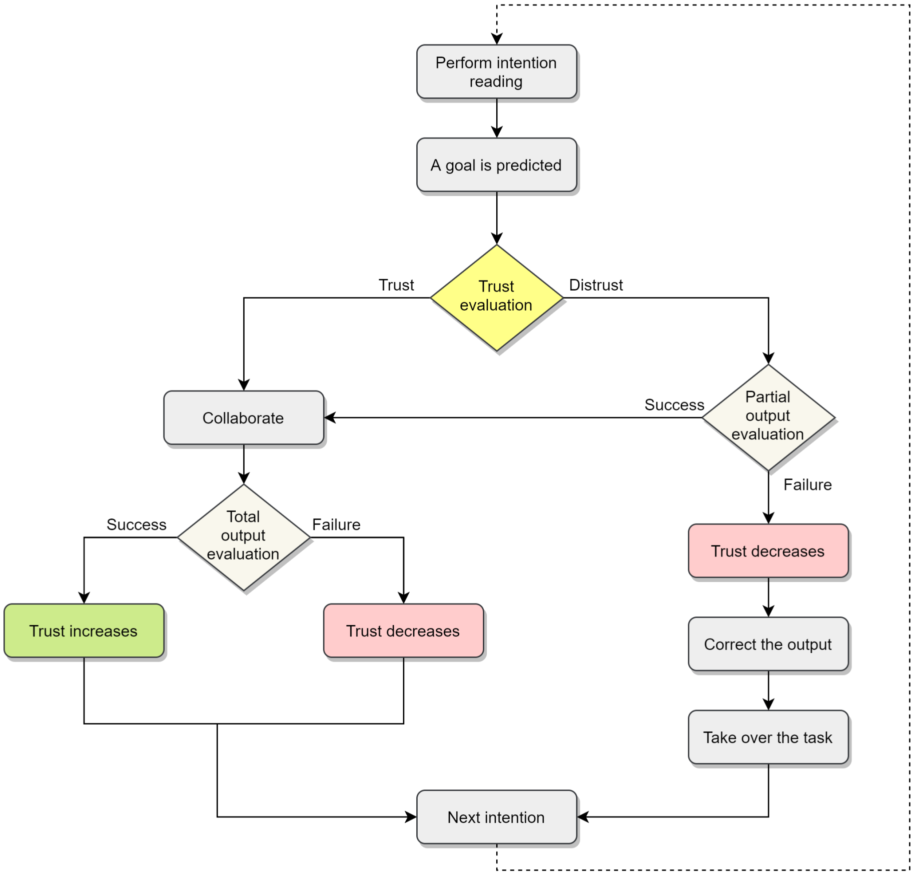
Now that we have described both the intention reading and the trust estimation models, we are going to focus on their integration with the purpose of achieving a cognitive architecture suitable for human-robot collaborations. As described in Figure 1, the main idea is that the trust model will act as a cognitive support for the intention reading, allowing the robot to fine-tune its behavior after having decided a general course of action. In particular, the robot will initially be trained on a set of goals and will thereafter try to understand which one is being pursued by its partner. Once a confident prediction is formulated, it will offer assistance in order to achieve the shared objective. The degree of help provided is influenced by the amount of trust the robot has in that specific person: if it thinks he or she have the appropriate knowledge or skills to complete the task, then it will act as an assistive peer, on the contrary it will start behaving more like a supervisor, observing more closely the partner, correcting their mistakes and, in general, assuming more of the responsibilities to ensure that the goal is eventually reached.

The workflow we are about to discuss is shown graphically in Figure 7. The interaction starts with the robot trying to identify the partner: in case of success it fetches their trust belief BN, otherwise it generates one on-the-fly through its episodic memory. The robot is assumed to naively trust the person that has trained it, so it will possess at least one BN in its memory. After this process is completed, it will start observing the human to read his or hers intention. Once a goal is predicted, the robot will perform a proactive trust evaluation in which it will ask itself if it expects the human to fail or succeed in the task at hand: this is done by setting X_R and Y_R as evidence and using the Message Passing algorithm to calculate the posterior probabilities for the rest of the network. At this point, the agent can use the probability distributions in nodes X_I and Y_I to infer the informant's behavior. Based on this evaluation, the robot will adopt one of two different approaches.

If the robot decides to trust the human, then it will collaborate towards the achievement of the predicted goal. Once the task is complete, it will judge the Total Output (TO) of the joint action: if the shared effort led to a successful, valid outcome its trust level towards the partner will increase, on the contrary it will decrease. If however the robot decides to distrust the partner, it will immediately inspect the Partial Output (PO), that is the portion of the task that has already been completed before an intention prediction was formulated. If the PO appears invalid, the robot will lower its trust level and will thereafter try to correct the mistake and take over the rest of the task. If instead the PO is a valid one, even if not the one which the robot had predicted, the agent will give the partner a chance to regain trust by collaborating and evaluating the TO, as described previously.

This workflow penalizes human partners who are both incapable or unwilling to contribute with an appropriate effort to the shared task, but at the same time gives distrusted people a chance to regain the trust of the robot. This is important, because failures could arise from temporary situations such as injuries or fatigue.

In an effort to include some features of Explainable AI (Hagras, 2018) into our system, the robot will try to be transparent and constantly communicate to its human partner any estimation results and any changes in its levels of trust. So, for example, if the robot doesn't trust the human to be able to accomplish a pursued goal, it will state that clearly, thus justifying its much more strict behavior. In particular, the robot will always state: the predicted goal, the estimated trust levels including any changes from trust to distrust or vice versa, its evaluation of the TO or PO and the explanation of why it believes that a task was unsuccessful. Finally, the agent will also try and justify its own errors: for example, if it realizes that the achieved goal was not the predicted one, but nevertheless was valid.



Supplemental Figure 7: The collaborative workflow. The robot reads the intention and decides if to trust or not its partner. In the former case, it provides assistance and evaluates the total output resulting from the collaboration, otherwise it adopts a more strict supervision on the human: if the partial output seems valid it gives them a chance to regain trust, otherwise it will take over the task and attempt to correct the mistakes. Related to Figure 2.

References

Aldrich, J. et al. (1997). Ra fisher and the making of maximum likelihood 1912-1922. *Statistical Science*, 12(3):162–176.

Bellman, R. (2013). *Dynamic programming*. Courier Corporation.

Cangelosi, A., Schlesinger, M., and Smith, L. B. (2015). *Developmental robotics: From babies to robots*. MIT Press.

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2016). Realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1611.08050*.

Cippitelli, E., Gasparrini, S., Gambi, E., and Spinsante, S. (2016). A human activity recognition system using skeleton data from RGBD sensors. *Computational intelligence and neuroscience*, 2016:21.

Douc, R. and Cappé, O. (2005). Comparison of resampling schemes for particle filtering. In *Image and Signal Processing and Analysis, 2005. ISPA 2005. Proceedings of the 4th International Symposium on*, pages 64–69. IEEE.

Erikson, E. H. (1993). *Childhood and Society*. W. W. Norton & Company.

Hagras, H. (2018). Toward human-understandable, explainable ai. *Computer*, 51(9):28–36.

Jha, S. and Busso, C. (2017). Probabilistic estimation of the driver's gaze from head orientation and position. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–6.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 157–224.

Malle, B. F., Moses, L. J., and Baldwin, D. A. (2001). *Intentions and intentionality: Foundations of social cognition*. MIT press.

Manzi, A., Dario, P., and Cavallo, F. (2017). A human activity recognition system based on dynamic clustering of skeleton data. *Sensors*, 17(5):1100.

Ojala, T., Pietikainen, M., and Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987.

Patacchiola, M. and Cangelosi, A. (2017). Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods. *Pattern Recognition*, 71:132 – 143.

Pelleg, D., Moore, A. W., et al. (2000). X-means: Extending k-means with efficient estimation of the number of clusters. In *ICML*, volume 1, pages 727–734.

Rekleitis, I. M. (2004). A particle filter tutorial for mobile robot localization. *Centre for Intelligent Machines, McGill University*, 3480.

Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28(5):675–691.

Vanderbilt, K. E., Liu, D., and Heyman, G. D. (2011). The development of distrust. *Child development*, 82(5):1372–1380.

Vinanzi, S., Cangelosi, A., and Goerick, C. (2020). In *2020 29th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).*

Vinanzi, S., Goerick, C., and Cangelosi, A. (2019). Mindreading for robots: Predicting intentions via dynamical clustering of human postures. In *2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 272–277.

Vinanzi, S., Patacchiola, M., Chella, A., and Cangelosi, A. (2019). Would a robot trust you? Developmental robotics model of trust and theory of mind. *Philosophical Transactions of the Royal Society of London B.*

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE.