Toward Robots' Behavioral Transparency of Temporal Difference Reinforcement Learning With a Human Teacher

Marco Matarese, Alessandra Sciutti, *Member, IEEE*, Francesco Rea, *Member, IEEE*, and Silvia Rossi, *Senior Member, IEEE*

***Abstract*—The high request for autonomous human–robot interaction (HRI), combined with the potential of machine learning (ML) techniques, allow us to deploy ML mechanisms in robot control. However, the use of ML can make robots' behavior unclear to the observer during the learning phase. Recently, transparency in HRI has been investigated to make such interactions more comprehensible. In this work, we propose a model to improve the transparency during reinforcement learning (RL) tasks for HRI scenarios: the model supports transparency by having the robot show nonverbal emotional-behavioral cues. Our model considered human feedback as the reward of the RL algorithm and it presents emotional-behavioral responses based on the progress of the robot learning. The model is managed only by the temporal-difference error. We tested the architecture in a teaching scenario with the iCub humanoid robot. The results highlight that when the robot expresses its emotional-behavioral response, the human teacher is able to understand its learning process better. Furthermore, people prefer to interact with an expressive robot as compared to a mechanical one. Movement-based signals proved to be more effective in revealing the internal state of the robot than facial expressions. In particular, gaze movements were effective in showing the robot's next intentions. In contrast, communicating uncertainty through robot movements sometimes led to action misinterpretation, highlighting the importance of balancing transparency and the legibility of the robot goal. We also found a reliable temporal window in which to register teachers' feedback that can be used by the robot as a reward.**

***Index Terms*—Human–robot interaction, humanoid robot, reinforcement learning (RL), social robotics, transparency.**

I. Introduction

THE more robots become autonomous and flexible, the more their behaviors need to be transparent. When interacting with complex intelligent artefacts, humans inevitably formulate expectations to understand and predict their behaviors. Indeed, robots' behaviors should be self-explanatory so that users can be confident in their knowledge of what these systems are doing and why [1]. Indeed, in the field, one of the interpretations of the term transparency is related to the *observability* or *predictability* of a system's behavior, and the possibility to infer its intentions, e.g., understanding what the robot is doing, why it is doing an action rather than another one, and what it is going to do next. Hence, it is also a mechanism that exposes an agent's decision-making process [1].

Moreover, the possibility of interpreting the behaviors of intelligent others, both in case of success and failures, is a fundamental characteristic of successful interactions: it affects human trust in automation [2]. This necessity is evident when dealing with machine learning (ML) algorithms that control robots' behaviors. Through ML techniques, robotic systems can understand and classify a large class of humans daily actions [3]; in turn, these actions are starting to be used in human–robot interaction (HRI) applications.

Among ML techniques, reinforcement learning (RL) is a powerful learning method that is widely used in robotics as it combines perception and decision-making [4]. RL agents make errors during their learning process not only because they have not yet acquired the necessary skills, but also because errors and exploration are intrinsically part of RL training processes. Hence, it is increasingly important to make the RL training transparent to human users. Broekens and Chetouani [5] also foresee that the transparency in robots' behaviors could improve the quality of the HRI, e.g., allowing the engagement of humans in complex interactive scenarios. Furthermore, they stated that transparency may have a direct impact on the robot learning process.

In this work, we refer to the term transparency by expanding the meaning of predictability and observability. Legibility enables observers to quickly infer an agent's objectives, while predictability refers to match what an observer would expect [6], [7]. In addition to those characteristics, the concept of transparency that we used also identifies the agent's will of intentionality in its communication. This presupposes a certain degree of autonomy

Manuscript received November 27, 2020; revised February 28, 2021 and July 16, 2021; accepted September 19, 2021. Date of publication October 19, 2021; date of current version November 12, 2021. This work was supported by the Starting Grant from the European Research Council under the European Union's Horizon 2020 research and innovation programme G.A. No 804388, wHiSPER. This article was recommended by Associate Editor J. Y. C. Chen. *(Corresponding author: Marco Matarese.)*

Marco Matarese is with the University of Genova and Italian Institute of Technology, 16163 Genova, Italy (e-mail: marco.matarese@iit.it).

Alessandra Sciutti and Francesco Rea are with the Italian Institute of Technology, 16163 Genova, Italy (e-mail: alessandra.sciutti@iit.it; francesco.rea@iit.it).

Silvia Rossi is with the University of Napoli Federico II, 80125 Napoli, Italy (e-mail: silrossi@unina.it).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/THMS.2021.3116119.

Digital Object Identifier 10.1109/THMS.2021.3116119

of the agent [8] that is not strictly required in contexts of legibility and predictability. Moreover, predictability does not necessarily require communicative intentionality from the agent.

We aiming at evaluating the effects of transparency on the quality of the subjective HRI [9] (e.g., the degree of well-being that depends on how both natural and complex the interaction is). The challenge we want to address is to make humans able to better understand robots' behaviors. In contrast, the majority of work in HRI has been concerned with enabling robots to understand human behaviors. Transparency allows people to anticipate others' behaviors [2] and be understood by others. Thus, transparency needs to be a mutual agreement between two (or more) partners. To build a mutual understanding between humans and robots [10], we need to ensure that robots' behaviors are always transparent.

An example of a robot assistant could give us insights about how robot transparency can improve the HRI. Let us imagine a disabled or elderly person asking to their robot assistant to go into the kitchen to take a delicate object on a high shelf. It could be useful to interact with a transparent and expressive agent because its uncertainty could implicitly communicate that it does not know what object the user is referring to or that it can not find it because it has been moved to another place.

In details, we focused on the transparency of the learning process. We designed a model starting from a temporal difference (TD) RL theory of emotion (ToE) developed by Broekens in [11]. His theory started from the definition of emotions as *"valenced reactions to (mental) events that modify future actions, grounded in bodily and homeostatic sensations"* [11]. This is founded upon both cognitive [12], [13] and biological-evolutionary theories [14], [15]. The TD-RL ToE states that TD error can model emotions as joy (distress) and hope (fear); thus, a robot—learning through a TD-RL based technique—can show its learning trend to the person with which it is interacting. Our model enables an agent to select the appropriate nonverbal behavior to both communicate its feelings related to its learning trend and to detect and interpret human feedback (which is used as a learning signal). Simulated emotions in robots are widely used because humans have the innate ability to interpret them and to use these signals to express their feelings [16]. Indeed, emotion expression is a language-independent and species-independent way to express the subjects' internal states [5]. As such, simulating emotions could be used to make learning robots more transparent to their human users and coworkers. Moreover, robots that display social behaviors are better liked by users and rated more positively than are robots that display neutral behaviors [17].

We conducted a user study to evaluate how an expressive behavior can make a robot more transparent to the human teacher while it is performing an RL task. Consequently, we wanted to make the robot's teacher more aware of its emotional state to build an easier HRI. Our aim was to investigate if managing the robot's emotional and behavioral expressivity via only the TD error (without training for the user) could make the HRI more efficient, likeable and effective. Since the robot had human teachers, our goal was also to improve the teaching interaction with no training for the user.

II. Background and Related Works

To improve the effectiveness of human–robot collaboration, the HRI community often relies on reproducing human–human interaction mechanisms. For many researchers, trust is the main reason to build transparent robots: some guidelines and models for HRI have been developed for this purpose [8], [18] [19].

Gaze cues can leak information about future robot intentions but they work better when they are produced by human-like robots [20], [21]. Gaze plays an important role in communicating information about the environment [22]. When people refer to objects in their environment, they are looking at those objects before naming or grasping them. Usually, people fixate upon objects one second or less before naming them [23]. Moreover, people are good at identifying the target of their partner's gaze and then using that information to predict their partner's next actions [24].

Behavioral transparency and robot's expressivity play crucial roles in HRI. Indeed, in [25], the authors pointed out that expressive robots are preferred to more efficient ones. In addition, Broekens and Chetouani [5] argued that researchers and designers must develop transparent learning protocols for robots and virtual agents. In this way, during online learning scenarios, a human teacher has the opportunity to read the current state of a learning agent intuitively and naturally. They further argue that such an intuitive signal could be a good basis for expressing the emotions on the TD error.

Transparency has been approached as the capability to answer simple queries about the control algorithms [26], as well as the capability to display information to support human–robot communication through visual interfaces [27]; it has also been considered the ability to show robots' inner states through emotions [25]. The main field in which robot transparency and predictability have been investigated is robot motion. This has occurred because it is one of the robots' features that is directly observable by the user [28]. Legible motion, which is planned to clearly express the robot's intent, leads to more fluent collaborations than does motion planned to match people's expectations [29]. Other approaches have dealt with improving transparency through the use of nonverbal cues [30]. Robot transparency in HRI can improve bidirectional communication between robots and human users: such a mutual communication is necessary for a natural interaction [31]. Indeed, the authors in [32], have shown that social models can improve learning tasks powered by RL frameworks. It is crucial that humans would naturally interpret robots' social behavior, especially if our objective is to improve the trustworthiness of robots: this has been proved to be possible through nonverbal cues [19].

The majority of approaches in the literature dealing with transparency use only explicit signals, such as gaze [33], but do not consider the role that emotions could have during the interaction. At this point, there is no generally accepted computational framework that links emotions to RL (see [34] for a review). Thus, robots and virtual characters that learn via RL currently lack the ability to produce and interpret emotions in line with their learning process. There are frameworks based on cognitive appraisal theory [35]–[37]. However, these models assume that

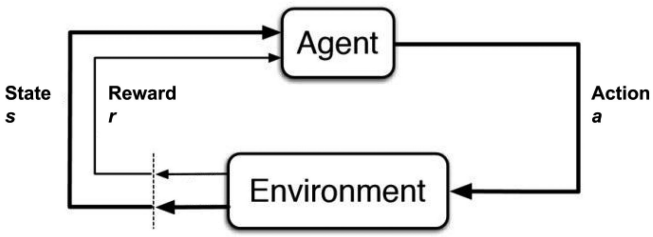


Fig. 1. RL general loop.

emotions arise from a cognitive reasoning process, not a learning one based on exploration, with positive and negative feedback. There are also works showing that robots and virtual agents can use human feedback as signals to influence their learning process based on RL [38]–[40]. In turn, this provides a good starting point for humans in the loop learning perspective.

As such, we wanted to investigate how people react to a robot expressing emotions during an RL task, and how this can improve robots' transparency during an online RL training. We aimed to investigate how transparent a robot behavior that is entirely driven by the TD is, i.e., that captures both the learning process and the current state, without participants' training. To enrich the existing literature, we considered emotions as not just social signals, but mostly as valuable reactions to learning actions [38]. To involve people more, we asked them to take the role of a teacher and to provide feedback to the robot, from which the robot could calculate a reward signal [41], [42]. In this manner, we wanted to make robots more transparent during their RL exploration process.

III. Methods

A. Reinforcement Learning

In the standard RL model (see Fig. 1), an agent is connected to an environment via both perception and action. The environment has a state $s \in S$, where S is the set of possible states, and the agent performs actions $a \in A(s_t)$, where $A(s_t)$ is the set of possible actions available in state s at time t [43]. At each step, the agent perceives the environment and selects which action to execute. This action changes the environment's state. The value of this state transition is communicated to the agent through a scalar reinforcement signal, r [44]. The agent's behavior is based on its policy π. In RL, this policy gets updated after each step taking into consideration the agent's experiences. In particular, the agent tries to maximise the total amount of the received rewards by maximising the sum of such rewards. Usually, a discount factor $\gamma \in (0, 1)$ is applied to every received reward to ensure there is a finite sum

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}. \tag{1}$$

The agent's goal is to find an optimal policy π_{opt} that maps every state with the best action the agent could do in that particular configuration of the environment.

B. Markov Decision Processes

Standard RL tries to resolve the problem of finding the optimal policy by satisfying a Markov process. The assumption (the Markov property) is that the probability distribution of the next state depends only on the previous pair $(\text{state}, \text{action})$. The probability to get into state s' coming from s after making the action a, $P^a_{s,s'}$, can be defined as follows:

$$P^a_{s,s'} = P(s_{t+1} = s' | s_t = s, a_t = a) \tag{2}$$

and the expected reward with respect to the same variables, $R^a_{s,s'}$, as follows:

$$R^a_{s,s'} = E(r_{t+1} | s_t = s, a_t = a, s_{s+1} = s'). \tag{3}$$

With $P^a_{s,s'}$ and $R^a_{s,s'}$ we can define the value $V^\pi(s)$ of each state s, for the policy π

$$V^\pi(s) = E_\pi(R_t | s_t = s) = E_\pi \left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s \right). \tag{4}$$

A state value is usually initialised and updated every time the state is visited. Since states are policy dependent, they can be used to update the policy. After each change of state, the values are updated as follows:

$$V_{k+1}(s) = \max_a \sum_{s'} P^a_{s,s'} \left[R^a_{s,s'} + \gamma V_k(s') \right]. \tag{5}$$

At this point, the policy becomes

$$\pi(s) = \arg\max_a \sum_{s'} P^a_{s,s'} \left[R^a_{s,s'} + \gamma V_k(s') \right]. \tag{6}$$

This algorithm can be used if the knowledge of the state space is complete, even though this does not always happen. TD learning updates those values after each visit. In its simplest method, it updates values as follows:

$$V(s_t) \longleftarrow V(s_t) + \alpha \left[r_{t+1} + \gamma V(s_{s+1}) - V(s_t) \right] \tag{7}$$

where $\alpha \in (0, 1)$ is a learning rate. After convergence, those values can be used to determine the best actions with an action selection method, e.g., with *softmax*.

C. TD-RL Theory of Emotions

As introduced in Section I, the TD-RL ToE developed in [5] is a key part of this study. Indeed, inspired by this ToE, we designed the robot's emotional-behavioral response we present here.

The TD-RL ToE was born because the authors in [5] recognised the lack of transparency of RL-based methods. Such a theory states that the value and reward functions are a good approximation of modeling RL-based emotions. TD-RL ToE is based on the assumption that all emotions are manifestations of a TD error [11]. This assumption is supported by the following two main arguments: 1) TD error and emotions' elicitation are similar, in the sense that both are feedback signals that result from evaluating a particular event, and 2) both TD error and emotions affect future behaviors.

In its formulation, TD-RL ToE defined emotions such as joy, distress, hope, and fear. Joy and distress have been thought of as

manifestations of positive or negative TD error [11], [38]. They refer to the present situation and they share the TD assessment with hope and fear, as this is one of their basic features [5]. However, hope and fear do not refer to the present; rather, they are about the future.

Our formulation started from the model cited above. In particular, we modeled emotions such as happiness and sadness, as well as behaviors like certainty and uncertainty. Just as in [5], we think that it is possible to make transparent learning robots through emotions simulation, especially if these emotions are related to the robots' learning process. In our opinion, a model based on the happiness-sadness dyad is very appropriate with consideration of the exploitation–exploration nature of RL techniques. However, we think that it could be more suitable to use a behavioral factor—such as (un)certainty—to implement the RL temporal dimension and the exploration–exploitation process typical of RL techniques.

D. RL Task

Our main research hypothesis was that during the learning process, endowing the robot with an emotional model of the behavior is sufficient to make the robot decision-making process more transparent to a human observer. We wanted to measure to what extent an emotional-behavioral response could make a robot more transparent while it was performing an RL task. To do so, we placed a humanoid robot (iCub) in a simple RL scenario: the robot had to learn a specific sequence of objects (coloured balls) that had been defined *a priori* by the investigator. The sequence was composed of five items, and it potentially had repetitions. To determine the sequence, the robot had to point to one ball after the other in the right sequence. In response to each action, the human teacher could give positive, neutral, or negative feedback. Since the sequence was composed of five objects, for each step of the learning process, a final state was reached after five actions. The experimental setup is shown in Fig. 2: the robot iCub was placed on a fixed metal support in front of a table. On this latter, there were five balls aligned in fixed positions as well as a joystick. The participants' chair was on the other side of the table.

E. RL Task Formulation

We needed an RL task that could be carried out in a real HRI scenario. For this purpose, because of its simplicity, we modeled the RL task described earlier as a Q-learning problem. We needed to have a simple space of states because the online learning process could not take a lot of time. Thus, the robot had five possible actions: pointing at each of the given balls. In addition, the state space was composed of five different states: each state represented which step in the sequence the robot was in. This gives the Q-Table a shape of 5×5.

The robot's multimodal behavior was selected according to the learning process and, in particular, it was selected to the following Bellman equation implementation [45]

$$Q(s,a) = Q(s,a) + \alpha \cdot (r(s) + \gamma \cdot \max_{a'} Q(s,' a') - Q(s,a)) \tag{8}$$

Fig. 2. Experimental setup.

where $r(s)$ is the reward given during the transition to state s that, in our case, was provided by the human teacher. This update rule uses the state and action of time t with the reward, state, and action on the time $t+1$ to compute the TD error. An update rule with this form is called associative because the learned values are associated with particular states of the environment. Thus, from the equation mentioned above, we extract the TD error as

$$\text{TD} = r(s) + \gamma \cdot \max_{a'} Q(s,' a') - Q(s,a). \tag{9}$$

TD error is a local estimate of the learning trend. Hence, the progress of the learning process can be measured with the TD error, which reflects the amount of change needed to the current estimate of the value of the state in which the agent is in [5]. Thus, it provides information about the correctness of the previously performed action in comparison with its estimation [46].

F. Robot Expressions

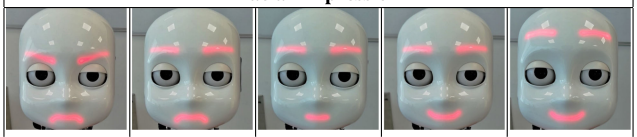
The TD-RL ToE postulates emotions as the manifestation of assessing the worsening or improvement of the current situation. In this direction, we defined the robot's emotional-behavioral response through three nonverbal communicative channels: the (un)certainty of the movement, the facial expression, and the gaze. Hence, we used each channel to display a particular feeling: the degree of confidence the robot has about the advancement of the learning process, the degree of satisfaction it has concerning the previous action's result, and an anticipatory signal for the next chosen action. All of these are communicative channels that are already used in HRI [20]. Moreover, we chose them with consideration of the learning task and the robot's physical capabilities.

1) (Un)certainty Movements: The (un)certainty of robot movements depended on the sign of the TD error as follows:

a) if TD < 0, then the robot's movement was uncertain;
b) if TD ≥ 0, then the robot's movement was certain.

Fig. 3. Uncertain movement's steps: (1) the rest position, from which all the robot movements began; (2) the thoughtful movement; (3) the final gaze; (4) the pointing action.

TABLE I
ROBOT'S FACIAL EXPRESSIONS AND TD ERROR VALUES

Facial Expression				
				
Very Sad	Sad	Neutral	Happy	Very Happy
TD Error				
-2	-1	0	+1	+2

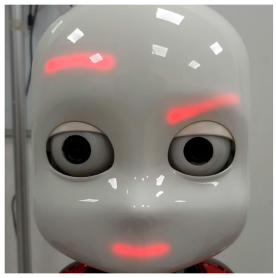


Fig. 4. Robot's doubtful expression.

During the experiment, the robot produced both certain and uncertain movements. We modeled the uncertainty through a fragmented movement of the robot arm that coincided with a gazing action. In particular, when the robot showed uncertain movements, before pointing its finger [see Fig. 3(4)], it moved its hand in mid-air to appear thoughtful and uncertain [see Fig. 3(2)]. Subsequently, it moved its head toward the chosen object [see Fig. 3(3)]. In contrast, we modeled certainty by having the robot perform a fast, well-directed pointing movement that was anticipated by a gazing action. Hence, certain actions were composed of two subactions: the anticipatory gaze and the proper pointing [see Fig. 3(3) and (4)].

2) Anticipatory Signal: The gaze followed the same rule as did the movements:

a) if TD < 0, then the robot's gaze was not anticipatory;
b) if TD ≥ 0, then the robot's gaze was anticipatory.

A gaze was considered "anticipatory" when it anticipated the following movement. For example, in our experiment, the robot looked at the chosen object before pointing at it. The aim of reproducing natural anticipatory gaze was to show the robot's focus of attention and consequently the degree of certainty in choosing the next object. In particular, people tend to look at a particular object right before grasping or using it. It has been shown that people can infer other people's intentions using just their gaze anticipatory movements [47] and these findings have already been used in both computer science and robotic fields [48], [49]. Such a signal has already been proved to be effective in increasing transparency by reducing uncertainty [33].

3) Facial Expressions: Moreover, the facial emotional reactions were modeled according to the TD error differently. During a preliminary study, we determined that the TD error values range was $[-2, +2]$. Hence, we defined five facial expressions to be associated with the integer values in that range. Table I summarises the association between the robot's facial expression and the TD error values. The particular facial expressions were chosen depending on the value closest to the TD error registered by the algorithm.

Here, we would like to emphasise that the robot behavior was not a just function of the human feedback. Rather, its behavior was modeled by the TD error; as such, it was a function of its current and previous states.

At the beginning of the experiments, iCub showed a neutral facial expression and uncertain behavior. Subsequently, the robot changed its facial expression immediately after the TD error was calculated: at the end of the pointing action. Thus, it maintained the selected expression for a few seconds; then, it became neutral.

For uncertain actions, during the uncertain submovements [see Fig. 3(2)–(4)], the robot changed its expression to one expressing doubt (see Fig. 4), and then it went neutral during the gaze and the pointing submovements until the calculation of the TD error and the consequent change of expression. As mentioned above, the robot's emotional-behavioral response was managed by the TD error; hence, also the (un)certainty of the robot as a function of such a value. By definition, the TD error is computed after the robot's action execution but, since we needed to model the robot's behavior at each step of its learning process, we stored the last TD error calculated for each state the robot has encountered. This way, when the robot was in a particular state s, the emotional-behavioral model used the TD error calculated in the same state s at the previous epoch. Once the action was completed, the TD was updated.

IV. User Study

A. Participants Feedback

Participants were required to give feedback to the robot during each action by using the joystick on the table. Participants gave continuous feedback to the robot's actions in the range $[-100, +100]$. Hence, it could be negative, neutral, or positive. We used this range because it gave us a good level of precision and granularity.

Starting with that feedback, we calculated the robot's reward with the following equation:

$$r = \frac{1}{D_a} \sum_{t=0}^{D_a} f_t \tag{10}$$

where D_a is the selected action duration (in hundredths of a second) and f_t was the feedback recorded during the tth joystick reading. The joystick reading frequency was 100 Hz. So f_t was the feedback read during the tth hundred of a second. From (10), it is clear that the reward was calculated as an average of the feedback recorded during the entire action duration.

Through a preliminary study, we found a state-action configuration for which 20 epochs was enough to converge. Thus, we provided a maximum of 20 epochs for each learning session. We set the Bellman meta-parameters [45] with standard values recommended in the literature. In particular, we had $\alpha = 0.8$ and $\gamma = 0.9$. For the actions selection strategy, we used an ϵ-decreasing method. As such, after each action, the probability to choose the best action—according to the Q-Table—increased. After every learning session, the Q-Table was cleaned. Therefore, during each interaction, the robot had to learn the right sequence from scratch.

B. Procedure

There were 23 participants, ($M_{\text{age}} = 27$, $\text{SD}_{\text{age}} = 8$), 12 of whom had been recruited from "Join the Science"[1] program. The participants recruited in this manner were university students or workers. The other 11 participants were Ph.D. students or external collaborators from the Italian Institute of Technology (IIT), Genova, Italy. These latter were from IIT's research units that do not work with robots or AI, thus, they had no experience with robots. The research protocol was approved by the Regional Ethical Committee (Comitato Etico Regione Liguria—Sezione 1). All participants provided their written informed consent and they received compensation of 15€.

The experiments consisted of two conditions: one *mechanical* and one *transparent*. In both conditions, the robot was required to learn the right sequence of balls taking into account the human feedback. Notably, in one session, the robot showed a mechanical behavior, while in the other one, it showed an emotional-behavioral response. We balanced the order of the sessions obtaining a counterbalanced within-subject user study.

When it was in the *mechanical* condition, the robot simply pointed at the balls to indicate the chosen sequence. The pointing action began from a rest position, as shown in Fig. 3(1). After performing the pointing gesture, it then returned to the rest position. In the mechanical condition, during the pointing action, the robot did not look at the selected object and it always showed a neutral facial expression (Table I, central figure). All movements produced by the robot were hard-coded. In particular, the robot's pointing movements were composed of the gesture of pointing toward the target and a maintenance phase, in which the robot kept its arm still. The duration of both was the same. From now on, with "pointing movement," we mean both the gesture toward the target and the maintenance. The kinematics of the different submovements, in both the *mechanical* and the *transparent* conditions, were the same. The duration of the robot actions varied depending upon whether the action was *mechanical*, certain, or uncertain. At the end of each action, the robot returned to the same rest position.

At the end of each condition, we asked the participants to respond to some questionnaires: the inclusion of other in the self (IOS) questionnaire [50], the Godspeed questionnaire [51], and the Mind Attribution Test [52]. We used these questionnaires to determine any perceived differences between the two robot behaviors.

C. Experimental Hypotheses

Our first hypothesis was that the proposed emotional-behavioral model could improve the legibility of the robot's behaviors. Hence, (H1) *a robot emotional and behavioral response, which is entirely driven by the TD error, is sufficient to make the robot's behaviors more transparent to the human teacher*. In particular, we tested H1 by studying the timing and the shape of participants' feedback in both behavioral conditions. Second, we expected that the robot's anticipatory signals would elicit reliable feedback. As such, (H2) *the recorded feedback following the robot gaze signal matched the final participants' reward*. To test this hypothesis, we checked the coherence between feedback recorded during both gaze and pointing movements; then, we performed statistical tests to reach reliable results. For our third hypothesis, we expected that (H3) *people should prefer to interact with an expressive robot compared with a mechanical one*. We tested H3 through questionnaires that participants answered after each interactive session with the robot.

V. Results

A. Learning Performance

We looked for differences in the robot's learning performance between the two types of sessions, regarding the number of epochs needed to reach the highest accuracy: *mechanical* (m), and *transparent* (t). We found that the learning performances were fairly similar regardless of the showed behavior (*two-tailed t-test* $t(17) = 1.1346$, $p = 0.2732$), with averages and the standard deviations of the number of epochs needed to reach the highest accuracy $\mu_m = 8.2$ with $\sigma_m = 3.5$, and $\mu_t = 9.2$ with $\sigma_t = 3.8$.

[1][Online]. Available: https://www.great-campus.it/join-the-science/

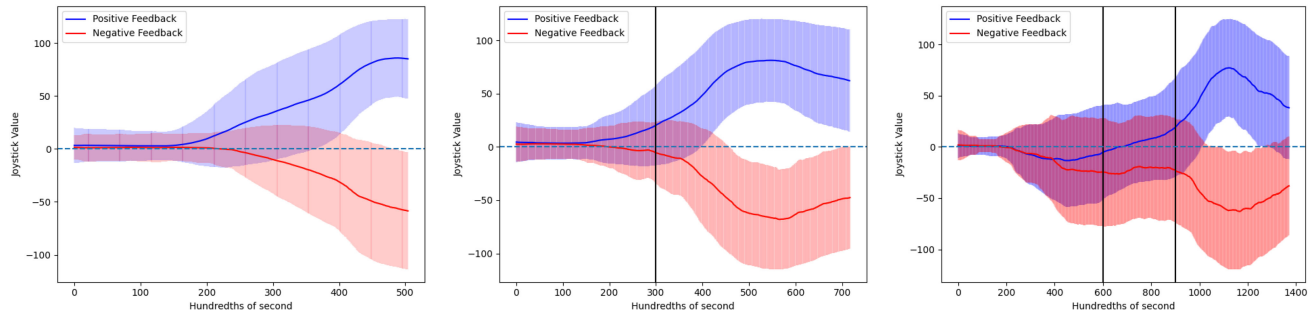


Fig. 5. Averages and standard deviations of participants' feedback given through the joystick over time. (a) *mechanical* movements, (b) *transparent* certain movements, and (c) *transparent* uncertain movements. The feedback values belonged to the range $[+100, -100]$. The vertical lines indicate when a robot subaction ends and another one begins (see description in the text).

B. Teacher Feedback Analysis

In this work, rather than focusing only on the robot's learning, we also focused on studying the participants' feedback. Our objective was to understand how the participants reacted to every type of robot's behavior.

To analyse the participants' joystick use, we split the data into six sets. A first division involved the robot's behavioral conditions. Thus, we separated the feedback given during the *mechanical* sessions from the feedback collected during the *transparent* ones. Subsequently, we split the feedback given during the *transparent* sessions depending on whether the participants provided this feedback to the robot's certain or uncertain actions. Finally, we divided the values of all these sets of feedback depending on whether they were negative or positive.

Fig. 5 shows the average participants' feedback in each type of session from the beginning of the robot's action sequence to its return to the rest position. Fig. 5(a) shows the results from the *mechanical* sessions, while Fig. 5(b) and (c) shows the results from the *transparent* ones: certain and uncertain actions, respectively. The vertical lines indicate the time in which a robot subaction ends and when another one begins. Therefore, in Fig. 5(a), we have no lines because the *mechanical* movements were composed of just the pointing movement. In contrast, in Fig. 5(b), we have one vertical line (third second), which indicates the moment in which the gazing submovement ends and when the pointing starts. Finally, in Fig. 5(c), we have two lines, the first one (sixth second) indicates the moment in which the uncertain submovement ends and the gazing begins, while the second line (ninth second) indicates the moment in which the gazing submovement ends and the following pointing starts. As we can see from the plots, we made a distinction between positive and negative feedback. We say that feedback is positive (in blue) if it is given to a correct robot action; in contrast, negative feedback (in red) is given to an incorrect robot action.

As one can see from the figure, during the *mechanical* sessions, the participants' feedback was concentrated at the end of the robot pointing gesture. This result shows that, during the *mechanical* sessions, participants understood the robot's intentions at the end of the robot's gesture toward the pointing target, just before the robot started keeping the pointing position still. In contrast, during the *transparent* sessions, participants were able to start giving their feedback during the robot's anticipatory gaze; thus, when the robot started pointing, they were already giving relevant feedback. One can also see that during the uncertain robot subaction [see Fig. 5(c) zeroth–sixth seconds], people gave negative feedback in both cases: very little for right actions and stronger negative feedback for wrong ones, on average (see Fig. 6). This type of action was not pointing or a signal communicating intentions. In the case of an uncertain movement [see Fig. 5(c) zeroth–sixth seconds], participants tended to immediately opt for negative feedback. Such feedback became even more pronounced when the robot pointed to the wrong ball [see Fig. 5(c) sixth–ninth seconds]. Conversely, more time is required to turn negative feedback into positive feedback, as a result of the robot gazing at the right target.

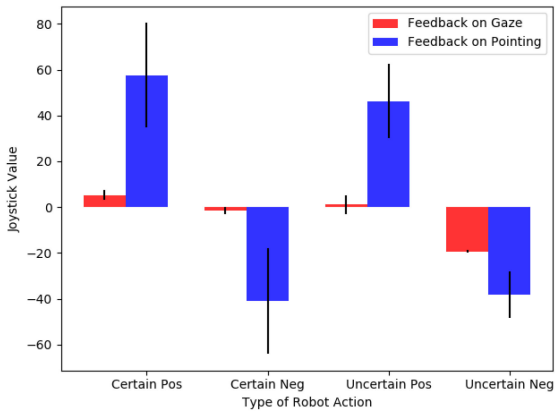


Fig. 6. Averages and standard deviations of feedback given during the gaze and the pointing movements.

From the plots in Fig. 5(b) and (c), it is clear that the anticipatory gaze was well perceived by the participants and that it had a fundamental role during the interaction, both in certain and in uncertain behaviors. During the latter, the feedback drastically changed with respect to the previous submovement. In addition, we can see from Fig. 5(b) and (c) that we had a decrease in the magnitude of participants' feedback at the end of the robot's pointing (especially for uncertain movements).

Fig. 6 shows the average feedback values provided during the robot gazing movement [see Fig. 5(b) zeroth–third seconds and Fig. 5(c) sixth–ninth seconds] as well as the ones provided during the pointing movement [see Fig. 5(b) from third second onwards and Fig. 5(c) from ninth second onwards]. The feedback collected during the gazing was consistent with those given during the following pointing action. Since we aimed to investigate the reliability of the teacher feedback in concomitance with every robot action, we further analyzed these data. Specifically, we separated the feedback depending on the type and correctness of the robot actions: thus, they were divided into either certain (right or wrong) or uncertain (right or wrong).

For each participant, we analyzed the average feedback recorded in the temporal windows in which the robot was performing the gazing and pointing actions (see Fig. 6). In particular, we performed *one-sample two-tailed t-tests* to determine whether the average feedback values were significantly different from zero: therefore, they could be considered reliable feedback, from which build reliable rewards. We considered feedback as "reliable" if it was clearly distinct from zero, showing a clear intent of participants to show either positive or negative feedback. During the certain robot behavior [see Fig. 5(b)], we found that, for correct robot actions, the feedback recorded during the robot gazing was significantly higher than zero ($\mu = 6.81, \sigma = 16.25$; *one sample two-tailed t-test* $t(22) = 12.098, p < 0.001$) (see Fig. 6). In contrast, for incorrect robot actions, the feedback was not significantly different from 0 ($\mu = 0.65, \sigma = 14.68$; *one sample two-tailed t-test* $t(22) = 0.785, p = 0.432$) (see Fig. 6). To identify a temporal window in which the feedback was significantly negative, we considered the feedback recorded from second 2.5 to the second 3.5 [see Fig. 5(b)], which was the feedback related to the end of the gazing movement and the beginning of the pointing one. In this time span, the feedback becomes significantly lower than zero ($\mu = -5.58$, $\sigma = 26.6$; *one sample two-tailed t-test* $t(22) = -3.682, p < 0.001$). In contrast, for uncertain robot actions, we found that the average feedback was significantly lower than zero during the uncertain submovement both during correct actions ($\mu = -4.86$, $\sigma = 17.9$; *one sample two-tailed t-test* $t(22) = -4.869, p < 0.001$) and incorrect robot actions ($\mu = -6.59, \sigma = 28.62$; *one sample two-tailed t-test* $t(22) = -5.521, p < 0.001$). Moreover, during the uncertain robot actions, the feedback registered during the gazing movement was significantly lower than zero for wrong actions ($\mu = -21.24$, $\sigma = 40.29$; *one sample two-tailed t-test* $t(22) = -7.649, p < 0.001$), but it did not significantly differ from zero for correct ones ($\mu = 1.41$, $\sigma = 35.2$; *one sample two-tailed t-test* $t(22) = 0.784, p = 0.469$). As we did for the previous analysis of certain movements, we slid the temporal windows of the latter condition. We found significant differences in the window starting between the seventh and the ninth seconds [see Fig. 5(c)]; thus, for the second half of the robot gazing movement, the average feedback became significantly larger than zero ($\mu = 4.69, \sigma = 34.38$; *one sample t-test* $t(22) = 2.399, p = 0.016$).

It has to be noted that, in both the *mechanical* and *transparent* sessions, even though we told participants that their feedback would be a continuous signal within the interval $[-100, +100]$, the participants tended to use the joystick as a discrete ± 1 signal. During the mechanical sessions, approximately 65% of the positive feedback fell into the range $[0, 25]$ and around 30% into the range $[75, 100]$. For certain negative feedback, we had similar results: approximately 59% belonged to the range $[0, -25]$ and around 33% to the range $[-75, -100]$. Finally, concerning uncertain feedback, approximately 62% of the positive feedback fell into the range $[0, 25]$ and 31% fell into the range $[75, 100]$. We had the same percentages for uncertain negative feedback but within the negative ranges $[0, -25]$ and $[-75, -100]$, respectively.

C. Teacher Wrong Feedback

We performed a total of 46 interactions. In 6 cases during *mechanical* sessions, and 5 cases during the *transparent* ones, the robot was not able to learn the entire sequence, only learning part of it. The reason for these failures can be attributed to the incorrect feedback provided by the participants and, consequently, to the way we computed rewards starting from those feedbacks. In fact, we found that the number of wrong feedbacks registered in sessions in which the robot learned the entire sequence was significantly lower than the number in those sessions in which the robot failed to learn the sequence (*two-tailed z-test* $z = -3.60$, $p < 0.001$). Thus, we investigated further. Depending on the environment's state, we split the robot actions into two sets: correct and incorrect actions. We considered a participant's feedback to be wrong when the robot's performed action was correct and the participant's feedback was negative and also when the robot performed the wrong action but received positive feedback; otherwise, we considered the feedback to be correct.

Subsequently, we assessed whether the frequency of wrong feedback was different in response to the two robot behaviors. To do so, we plotted the frequency of wrong feedback for all the subjects when they were confronted with the *mechanical* behavior against the frequency exhibited of wrong feedback by the same subjects when facing the *transparent* robot (see Fig. 7). From that plot, it emerges that the frequency of errors was similar during the two conditions, with a consistent behavior for a given subject in both (i.e., the participants that made more mistakes did so irrespective of the robot's condition). However, on average, we registered a nonsignificant higher frequency of wrong feedback during the *transparent* sessions (see Fig. 7) ($\mu_m = 10.28$, $\sigma_m = 10.21$ and $\mu_t = 14.58$, $\sigma_t = 12.28$).

The main reason participants provided wrong feedback was found to be attributed to the way we computed the rewards: considering the joystick values through the whole duration of the action. These results suggest that the uncertain movement sometimes was a threat to robot transparency and that, if we considered the feedback starting from the robot gaze, we could have gotten fewer errors and better performances.

D. Questionnaires

We used paired t-tests (with a confidence interval of 95%) to determine differences between the averages of the questionnaires' answers as they relate to the *transparent* and *mechanical*

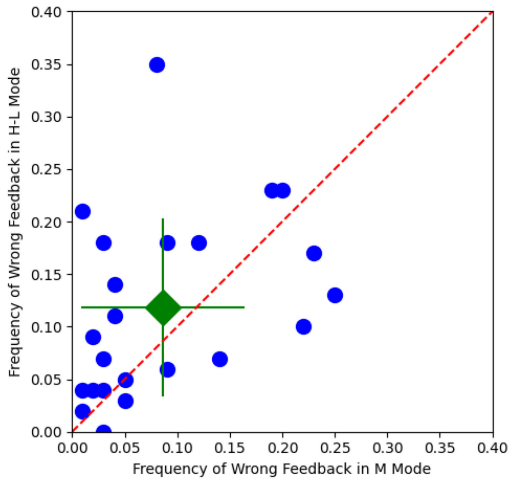


Fig. 7. Wrong feedback frequencies registered in both mechanical and transparent sessions. Average values and std errors are also indicated.

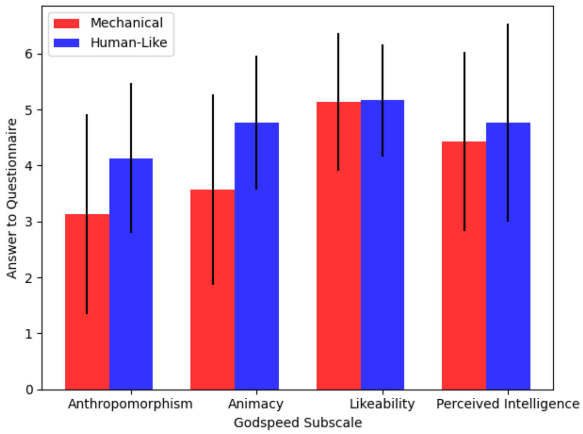


Fig. 8. Averages and standard deviations of the answers given to the Godspeed test.

conditions. The results of the mind attribution test showed significant differences in the robot's ability to feel pleasure ($t(22) = -3.49$, $p = 0.002$) and joy ($t(22) = -4.54$, $p < 0.001$). This means that during the *transparent* sessions, people perceived the robot to be more able to feel emotions during its learning process. In addition, there were differences regarding the robot's ability to plan its action ($t(22) = -2.55$, $p = 0.018$), to have self-control ($t(22) = -2.61$, $p = 0.015$), and to recognise emotions ($t(22) = -2.79$, $p = 0.01$); participants perceived their role to be more important during the *transparent* conditions than in the mechanical sessions.

The Godspeed test provides a scale that goes from 0 to 7. In this test, participants provided higher values for the *transparent* sessions than they did for the *mechanical* ones. There were significant differences between the robot's anthropomorphism ($t(22) = 2.158$, $p = 0.036$) and animacy ($t(22) = 2.765$, $p = 0.008$). We have reported the means and standard deviation values of the Godspeed test in Fig. 8. We also found a significant difference in answers given on the IOS questionnaire ($t(22) = -3.54$, $p = 0.001$): people felt closer to the robot during the *transparent* sessions than they did during the *mechanical* ones.

Through open-ended questions, approximately 14% of the participants explicitly complained about some inconsistencies in the robot's facial expressions. As stated earlier, the robot reactions were managed by the TD error while people expected more of a match between their positive/negative feedback and the robot's happy/sad expressions.

The open-ended questions also revealed that about 50% of the participants explicitly claimed the iCub's more expressive behavior was able to put forth the idea that the robot cared about the task. Moreover, the participants provided more feedback in the *transparent* sessions than they did in the *mechanical* sessions. People perceived indecision and the need for more help when iCub showed insecure behaviors.

VI. Discussion

We found no significant differences in the learning performance, regarding the number of epochs needed to reach the highest accuracy, between the two behavioral conditions of the robot. This result was not particularly surprising: we expected the robot to receive similar reward values in the two conditions. Rather, the differences we were looking to find concerned the timing of participants' feedback. Specifically, how the robot's different behaviors would change the teachers' responses.

A. Robot Expressiveness

In general, our results confirmed the fundamental role of robot gaze in improving transparency in HRI. Moreover, in the case of uncertain actions, we supposed that the uncertainty submovement could be correctly interpreted by the subjects as an index of the exploration process of the RL algorithm and, consequently, to an action that is probably wrong. Indeed, such negative feedback was corrected as soon as the anticipatory gaze was performed. This also suggests that the *transparent* expressions of uncertainty can "pause" an observer, much like what happens in human–human interactions. We noted that expressing doubt or uncertainty with the robot's movements can lead to misinterpretation and have a bad effect on the robot's transparency. The choice to reveal the robot's uncertainty through movement segmentation had a double effect: on the one hand, it triggered earlier feedback, though this was always negative, which might be seen as a sign of increased participation and engagement by human participants. On the other hand, it puzzled participants, who sometimes suggested that they were confused by this behavior, making them unable to infer the robot's goal. As a consequence, it could be relevant to consider different approaches to show uncertainty without compromising the legibility of the movement itself [10]. Not all the expressive signals used by the robot in the *transparent* condition had the same positive impact on transparency. Implementing basic sensorimotor regularities proper of human behavior, such as the anticipatory gazing toward the target of a future action [53], sped up participants' understanding of actions. Conversely, mapping the robot facial expressions to the degree of satisfaction based on the previous actions' effect proved to be less effective. This could

be because participants considered robot facial expressions to be an explicitly communicative social cue and, therefore, they expected the robot to respond to the social interaction rather than to the robot's internal evaluation. Indeed, in the context of human–human interaction, people often set their facial expressions voluntarily for a social purpose, even this is in contrast to their hidden feelings (e.g., non-Duchenne or "social" smile). Hence, participants expected a simpler coupling between the robot's facial expressions and the feedback they provided. This suggests that participants perceived such a communicative cue as a high-level signal. As such, these results suggest that the robot's facial expressions might be better suited to provide explicit feedback to the partner's actions rather than supporting transparency about its internal states. Thus, we should model them simply starting from the received reward.

We aimed to model the robot's transparency to be naturally perceived by the participants and entirely guided by the TD error. For this purpose, we needed signals that could be intuitively understood because they derive from our interaction vocabulary. We can not exclude that a part of the observed benefits on transparency was related to the effect of interacting with a robot behaving in a human-like manner.

Our results show that people prefer to interact with an expressive robot also in learning scenarios. Moreover, the TD error is a good estimation of the learning trend and our emotional-behavioral response resulted to be coherent with this latter, even though some adjustments are needed: some communicative signals (e.g., the facial expressions) could be better modeled through a simpler reward-based approach, and the way by which the robot express uncertainty (a fragmented movement) should be fine-tuned. In our opinion, our work provides a good starting point in the investigation of TD-based models for robot transparency during RL tasks.

B. Participants Feedback

During the *transparent* sessions, the user feedback started decisively during the gaze, meaning that this occurred before the pointing movement would begin. Thus, we can state that the anticipatory gaze was enough for the human participants to understand the robot's intentions. This confirms our hypothesis H1. On average, the feedback collected after the anticipatory gaze signal was consistent with the feedback that was given during the robot pointing action. However, the two intensities were very different from each other. For this reason, additional statistical analysis was needed to find the temporal window from which we could individuate reliable feedback. We found that, for some of our experimental conditions, reliable feedback began in concomitance with the beginning of the gazing movement. Hence, we can claim that our hypothesis H2 was partially confirmed. It cannot be considered fully confirmed because we also found that the feedback given during the gazing movement was not reliable enough in all our experimental conditions. In particular, in two cases (i.e., certain wrong actions and uncertain correct actions), we found reliable feedback sliding to the right the temporal window of reference. However, just a little slide was needed to find reliable feedback: in fact, at least half of this feedback belonged to the gazing temporal window. A possible explanation for this effect could be that, especially with regard to uncertain robot actions, the feedback recorded during the gazing movement was strongly influenced by the previous movement by the robot. This occurred because of the continuous nature of the teachers' feedback. In the latter scenarios, the feedback signal was not reliable enough. However, the closer we moved toward the pointing temporal window, the more reliable the teacher feedback became. From these results, we can certainly state that participants perceived the robot gaze as very informative regarding its intentions. Hence, we could start recording the teacher's feedback during this stage, but not before it. In addition, we should pay particular attention to robots expressing uncertainty and we should start recording the teacher's feedback only when we are sure about its reliability: this occurs in the middle of the anticipatory robot movement.

Therefore, we can exploit the information on feedback timing to compute the rewards in a more precise way: we could start reading the feedback from the moment the anticipatory signals start, taking the maximum value registered in the time window starting from the anticipatory signal and ending at the time half the action is executed. Since the participants used the joystick as a ± 1 signal, we could also approximate the reward received depending on its sign. Such an anticipated reward could be exploited during the learning process to block the execution of wrong actions. To do this, we would need to enrich the interaction by introducing social cues that could explain that the robot understood its failure. This could make the interaction more natural since, just as when we are teaching something to children, we would not expect them to persist in wrong actions after the moment we tell them they are totally mistaken.

In both *transparent* types of movements, we recorded a decrease in the magnitude of participants' feedback at the end of the robot's pointing. This phenomenon is due to a return in a neutral 0 position of the joystick. In particular, the majority of participants gave the maximum feedback (in magnitude) and then they returned back to the neutral position. This phenomenon was more emphasised during the human-like sessions because 1) they had more time to provide their feedback and 2) they understood the robot's intention earlier.

C. Questionnaires

The results of the questionnaires showed that the participants noted the differences between the two robot's behaviors and that they appreciated more the *transparent* one. Moreover, they felt closer to the robot during the *transparent* sessions. These results confirmed our hypothesis H3. Even though the *transparent* behavior made the robot more transparent, the more responsive the robot was, the more people expected that it would be able to have a complex interaction. Almost all the participants noted the robot's lack of speech communication; this was especially the case during those times they would have liked to ask the robot for explanations, particularly after repeated robot failures.

D. Limitations and Future Works

The findings derived from this work could be generalised to more complex tasks. However, to this aim, it would be necessary to overcome a few limitations of the current settings.

First of all, we need to resolve the simplifications introduced in our experiments. Participants' use of the joystick suggests a change in the method by which they provide the feedback: from a joystick to a reading of the teacher status via a valence-arousal-based classification. Reading both teachers' reactions and nonverbal cues could allow having a good approximation of reward signals. Both the anticipatory signals and the expression of uncertainty could allow us to redefine the online training phase making the robot able to suspend part of it depending on the teacher's reactions. This could redefine the RL paradigm to better align it to a more natural interaction from the human teacher's point of view. Moreover, issues related to RL tasks' complexity may require a two-steps learning strategy: training via simulation, and tuning and customisation via online training as in our experiments.

Finally, in modeling the robot's behavior, we used the gazing movement only when the robot was certain about the action it was going to perform. However, due to the important role that such signal has played, we plan to investigate further the impact of the robot's gaze in transparency by standardising it in both conditions. In particular, we plan to model TD error in one condition on a happy-sad scale, and on an uncertain-certain scale in a second condition.

VII. Conclusion

In this work, we studied robot transparency during RL tasks by providing a TD-based emotional-behavioral model to a robot (iCub). For this purpose, we designed a user study in which a robot performs an RL task, while a human teacher provides it feedback through a joystick. We studied differences in participants' reactions during two experimental conditions: they had to interact with a *transparent* and a *mechanical* robot.

Our model provided for the management of the robot's behavior by the TD-error to obtain a coherent and solid expressivity. However, this turned out to be an approximation that did not produce optimal results. Even though our multimodal behavioral-emotional response had a good effect on transparency, we need to investigate further how to differentiate signals related to social aspects and those related to the robot's internal state. Users without training expect that the robot's emotions would reflect the received feedback rather than the learning trend. The approach we used resulted to be effective in transparency more from the behavioral side than from the emotional one, but not always: gaze movements were crucial in understanding the robot's intentions, while uncertain movements sometimes led to misinterpretation. Nevertheless, our model made the HRI more likeable: teachers perceived the transparent robot as more involved in its task. Moreover, we do not exclude that this latter was due to a robot using communicative signals that were familiar to humans.

Our experiment provided for a simple RL problem because it had to be faced in an experimental context. Real RL problems are not so easy, thus usually they require a learning phase done in simulation. We think it could be very useful—after the learning phase—to have an interactive customisation phase in which the system can be fine-tuned based on the user's needs. In this scenario, interacting with a transparent robot can bring several benefits for both the HRI and learning performance, e.g., it could give clues about errors of bias the robot faced during the learning phase, or the robot could exclude learning paths that the teacher considers useless.

We need to investigate further what is the role of social signals in robots' transparency. Simply adding information is not enough. We need to study deeper how humans interact with each other; in particular, how humans differentiate communicative signals to assess their social objectives. This way, we could try to reproduce the interaction mechanisms we use in our everyday lives to improve the HRI and to make robots' behaviors more natural for their human partners.

References

[1] R. Wortham, A. Theodorou, and J. Bryson, "What does the robot think? transparency as a fundamental design requirement for intelligent systems," in *Proc. IJCAI Workshop Ethics Artif. Intell.*, Jun. 2016.

[2] J. Iden, "Belief, judgment, transparency, trust: Reasoning about potential pitfalls in interacting with artificial autonomous entities," in *Proc. Robot. Sci. Syst. 2017 Workshop/Morality Soc. Trust Auton. Robots*, Eds., N. Amato, S. Srinivasa, and N. Ayanian, Cambridge, MA, Aug. 2017.

[3] L. Raggioli and S. Rossi, "A reinforcement-learning approach for adaptive and comfortable assistive robot monitoring behavior," in *Proc. 28th IEEE Int. Conf. Robot Hum. Interactive Commun.*, 2019, pp. 1–6.

[4] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1238–1274, 2013.

[5] J. Broekens and M. Chetouani, "Towards transparent robot learning through TDRL-based emotional expressions," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 352–362, Apr.–Jun. 2021.

[6] A. D. Dragan, K. C. Lee, and S. S. Srinivasa, "Legibility and predictability of robot motion," in *Proc. 8th ACM/IEEE Int. Conf. Hum.-Robot Interact.*, 2013, pp. 301–308.

[7] Y. Zhang, S. Sreedharan, A. Kulkarni, T. Chakraborti, H. H. Zhuo, and S. Kambhampati, "Plan explicability and predictability for robot task planning," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2017, pp. 1313–1320.

[8] J. B. Lyons, "Being transparent about transparency: A model for human-robot interaction," in *Proc. AAAI Spring Symp. Ser.*, 2013, pp. 48–53.

[9] C. S. Wendt and G. Berg, "Nonverbal humor as a new dimension of HRI," in *Proc. RO-MAN 18th IEEE Int. Symp. Robot Hum. Interactive Commun.*, 2009, pp. 183–188.

[10] A. Sciutti, M. Mara, V. Tagliasco, and G. Sandini, "Humanizing human-robot interaction: On the importance of mutual understanding," *IEEE Technol. Soc. Mag.*, vol. 37, no. 1, pp. 22–29, Mar. 2018.

[11] J. Broekens and L. Dai, "A TDRL model for the emotion of regret," in *Proc. 8th Int. Conf. Affective Computing Intell. Interaction (ACII)*, 2019, pp. 150–156, doi: 10.1109/ACII.2019.8925441.

[12] A. Ortony, G. L. Clore, and A. Collins, *The Cognitive Structure of Emotions*. Cambridge, U.K.: Cambridge Univ. Press, 1990.

[13] A. Moors, P. C. Ellsworth, K. R. Scherer, and N. H. Frijda, "Appraisal theories of emotion: State of the art and future development," *Emotion Rev.*, vol. 5, no. 2, pp. 119–124, 2013.

[14] J. Panksepp, *Affective Neuroscience: The Foundations of Human and Animal Emotions*. London, U.K.: Oxford Univ. Press, 1998.

[15] K. R. Scherer, A. Schorr, and T. Johnstone, Eds., *Appraisal Processes in Emotion: Theory, Methods, Research*. London, U.K.: Oxford Univ. Press, 2001.

[16] F. Cavallo, F. Semeraro, L. Fiorini, G. Magyar, P. Sinčák, and P. Dario, "Emotion modelling for social robotics applications: A review," *J. Bionic Eng.*, vol. 15, no. 2, pp. 185–203, 2018.

[17] S. Rossi, M. Staffa, and A. Tamburro, "Socially assistive robot for providing recommendations: Comparing a humanoid robot with a mobile application," *Int. J. Social Robot.*, vol. 10, no. 2, pp. 265–278, Jan. 2018.

[18] R. H. Wortham and A. Theodorou, "Robot transparency, trust and utility," *Connection Sci.*, vol. 29, no. 3, pp. 242–248, 2017.

[19] J. J. Lee, B. Knox, J. Baumann, C. Breazeal, and D. DeSteno, "Computationally modeling interpersonal trust," *Front. Psychol.*, vol. 4, pp. 893–907, 2013.

[20] B. Mutlu, F. Yamaoka, T. Kanda, H. Ishiguro, and N. Hagita, "Nonverbal leakage in robots: Communication of intentions through seemingly unintentional behavior," in *Proc. 4th ACM/IEEE Int. Conf. Hum. Robot Interact*., 2009, pp. 69–76.

[21] H. Admoni, C. Bank, J. Tan, M. Toneva, and B. Scassellati, "Robot gaze does not reflexively cue human attention," in *Proc. Annu. Meeting Cogn. Sci. Soc.*, vol. 33, no. 33, 2011.

[22] H. Admoni and B. Scassellati, "Social eye gaze in human-robot interaction: A review," *J. Hum.-Robot Interact.*, vol. 6, no. 1, pp. 25–63, 2017.

[23] C. Yu, P. Schermerhorn, and M. Scheutz, "Adaptive eye gaze patterns in interactions with human and artificial agents," *ACM Trans. Interactive Intell. Syst.*, vol. 1, no. 2, pp. 1–25, 2012.

[24] J.-D. Boucher *et al.*, "I reach faster when i see you look: Gaze effects in human-human and human-robot face-to-face cooperation," *Front. Neurorobot.*, vol. 6, pp. 3–14, 2012.

[25] A. Hamacher, N. Bianchi-Berthouze, A. G. Pipe, and K. Eder, "Believing in bert: Using expressive communication to enhance trust and counteract operational error in physical human-robot interaction," in *Proc. 25th IEEE Int. Symp. Robot Hum. Interactive Commun.*, 2016, pp. 493–500.

[26] B. Hayes and J. A. Shah, "Improving robot controller transparency through autonomous policy explanation," in *Proc. Int. Conf. Hum.-Robot Interact.*, 2017, pp. 303–312.

[27] A. R. Selkowitz, C. A. Larios, S. G. Lakhmani, and J. Y. Chen, "Displaying information to support transparency for autonomous platforms," in *Proc. Adv. Hum. Factors Robots Unmanned Syst*., 2017, pp. 161–173.

[28] B. Busch, J. Grizou, M. Lopes, and F. Stulp, "Learning legible motion from human-robot interactions," *Int. J. Social Robot.*, vol. 9, no. 5, pp. 765–779, 2017.

[29] A. D. Dragan, S. Bauman, J. Forlizzi, and S. S. Srinivasa, "Effects of robot motion on human-robot collaboration," in *Proc. Int. Conf. Hum.-Robot Interact.*, 2015, pp. 51–58.

[30] C. Breazeal, C. D. Kidd, A. L. Thomaz, G. Hoffman, and M. Berlin, "Effects of nonverbal communication on efficiency and robustness in human-robot teamwork," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2005, pp. 708–713.

[31] A.-L. Vollmer *et al.*, "Robots show us how to teach them: Feedback from robots shapes tutoring behavior during action learning," *PLoS One*, vol. 9, no. 3, 2014, Art. no. e91349.

[32] A. Najar, O. Sigaud, and M. Chetouani, "Social-task learning for HRI," in *Proc. Int. Conf. Social Robot.*, 2015, pp. 472–481.

[33] A. L. Thomaz and C. Breazeal, "Teachable characters: User studies, design principles, and learning performance," in *Proc. Int. Workshop Intell. Virtual Agents*, 2006, pp. 395–406.

[34] T. M. Moerland, J. Broekens, and C. M. Jonker, "Emotion in reinforcement learning agents and robots: A survey," *Mach. Learn.*, vol. 107, no. 2, pp. 443–480, 2018.

[35] J. Dias, S. Mascarenhas, and A. Paiva, *FAtiMA Modular: Towards an Agent Architecture With a Generic Appraisal Framework*. Cham, Switzerland: Springer International Publishing, 2014, pp. 44–56.

[36] S. Marsella *et al.*, "Computational models of emotion," *A Blueprint Affect. Comput.-A Sourcebook Manual*, vol. 11, no. 1, pp. 21–46, 2010.

[37] S. Marsella and J. Gratch, "EMA: A model of emotional dynamics," *Cogn. Syst. Res.*, vol. 10, no. 1, pp. 70–90, 2009.

[38] J. Broekens, "Emotion and reinforcement: Affective facial expressions facilitate robot learning," in *Proc. Artif. Intell. Hum. Comput*., 2007, pp. 113–132.

[39] W. B. Knox and P. Stone, "Combining manual feedback with subsequent MDP reward signals for reinforcement learning," in *Proc. 9th Int. Conf. Auton. Agents Multiagent Syst.*, 2010, pp. 5–12.

[40] A. L. Thomaz *et al.*, "Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance," in *Proc. AAAI Conf. Artif. Intell.*, 2006, vol. 6, pp. 1000–1005.

[41] W. B. Knox, B. D. Glass, B. C. Love, W. T. Maddox, and P. Stone, "How humans teach agents," *Int. J. Social Robot.*, vol. 4, no. 4, pp. 409–421, 2012.

[42] W. B. Knox, P. Stone, and C. Breazeal, "Training a robot via human feedback: A case study," in *Proc. Int. Conf. Social Robot*., 2013, pp. 460–470.

[43] R. S. Sutton *et al.*, *Introduction to Reinforcement Learning*, vol. 135. Cambridge, MA, USA: MIT Press, 1998.

[44] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *J. Artif. Intell. Res.*, vol. 4, pp. 237–285, 1996.

[45] R. Bellman, "Dynamic programming," *Science*, vol. 153, no. 3731, pp. 34–37, 1966.

[46] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Mach. Learn.*, vol. 3, no. 1, pp. 9–44, 1988.

[47] S.-J. Blakemore and J. Decety, "From the perception of action to the understanding of intention," *Nature Rev. Neurosci.*, vol. 2, no. 8, pp. 561–567, 2001.

[48] A. Doshi and M. M. Trivedi, "On the roles of eye gaze and head dynamics in predicting driver's intent to change lanes," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 3, pp. 453–462, Sep. 2009.

[49] W. Yi and D. Ballard, "Recognizing behavior in hand-eye coordination patterns," *Int. J. Humanoid Robot.*, vol. 6, no. 3, pp. 337–359, 2009.

[50] A. Aron, E. N. Aron, and D. Smollan, "Inclusion of other in the self scale and the structure of interpersonal closeness," *J. Pers. Social Psychol.*, vol. 63, no. 4, pp. 596–612, 1992.

[51] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *Int. J. Social Robot.*, vol. 1, no. 1, pp. 71–81, 2009.

[52] F. Ferrari, M. P. Paladino, and J. Jetten, "Blurring human-machine distinctions: Anthropomorphic appearance in social robots as a threat to human distinctiveness," *Int. J. Social Robot.*, vol. 8, no. 2, pp. 287–302, 2016.

[53] R. S. Johansson, G. Westling, A. Bäckström, and J. R. Flanagan, "Eye-hand coordination in object manipulation," *J. Neurosci.*, vol. 21, no. 17, pp. 6917–6932, 2001.