Unveiling the Learning Curve: Enhancing Transparency in Robot's Learning with Inner Speech and Emotions

Georgios Angelopoulos[1], Carmine Di Martino[2], Alessandra Rossi[1,2], and Silvia Rossi[1,2]

***Abstract*— The lack of transparency in robotic learning processes poses a significant challenge to effective human-robot collaboration. This is particularly relevant in non-industrial settings because it prevents humans from adequately comprehending a robot's intentions, progress, and decision-making rationale, which is essential for seamless interaction. To address this issue, this work presents a study where users observe a robot endowed with three distinct emotional/behavioural mechanisms for conveying transparent information about its learning process. The proposed mechanisms use inner speech, emotions, and a combination of the two communication styles (hybrid). To assess and evaluate the transparency of these behavioural models, a between-subject study was conducted with 108 participants. Results indicate that the people's perception of the robot's warmth dimension increased when it utilized a hybrid model to explain its learning state. Additionally, increased transparency was observed when the robot used inner speech during the learning process.**

I. Introduction

In human-centric environments, robots must acquire new skills and tasks on-site, as it is impractical for designers to pre-embed all the necessary knowledge to complete their tasks. Consequently, while learning, robots must exhibit transparent behaviours to maintain an effective human-robot interaction (HRI), and not lose trust. Indeed, transparency allows them to gauge the extent to which a robot has mastered a task before collaborating with it [1].

Moreover, it is essential that a robot conveys its intended actions transparently also to minimize uncertainty. As humans possess a natural ability to behave and infer others' behaviour, we believe it is essential to study these interactions to create robots that exhibit transparent behaviours [2]. Key aspects of human behaviours used to foster natural communication, and therefore transparency, include various indirect and unconscious signals, often involving a combination of modalities such as gaze, facial expressions, emotions, and posture [3]. In HRI, it has also been observed that a stereotypical motion (a standardized movement) [4] and the use of familiar cues [5] could make the robot's behaviours more legible and predictable.

This work has been supported by the European Union's Horizon 2020 research and innovation programme under the MSCA PERSEO No 955778, Italian PON R&I 2014-2020 - REACT-EU (CUP E65F21002920003), and by the Italian MUR and EU under the CHIST-ERA IV project COHERENT (PCI2020-120718-2).

[1]G. Angelopoulos is with the Interdepartmental Center for Advances in Robotic Surgery - ICAROS, University of Naples Federico II, Naples, Italy georgios.angelopoulos@unina.it

[2]C. Di Martino, A. Rossi, and S. Rossi are with the Department of Electrical Engineering and Information Technologies - DIETI, University of Naples Federico II, Naples, Italy {alessandra.rossi, silvia.rossi}@unina.it

Another social cue used to communicate with humans is through emotional expressions. Affective facial cues in robots can enhance transparency in human-robot interactions by resulting in an essential modality to understand the robot's internal state and decision-making processes [6], [7]. When a robot conveys emotions, such as happiness or frustration, humans can better interpret the robot's intentions and actions, making its behaviour more predictable and accessible. Furthermore, emotionally expressive robots can adapt their responses based on the human partner's emotional state, so facilitating collaborative and supportive interactions [8]. However, it is important to consider the potential downsides of emotional expression in robots. For example, emotions may be misinterpreted as social feedback, leading to confusion or miscommunication, and it could inadvertently influence humans in unintended ways [7].

Providing verbal explanations is another way to improve transparency in HRI [9]. When the robot is learning a task alone, without interacting with the user, the concept of inner speech [10] can become a crucial element. More specifically, inner speech is a covert inner monologue or dialogue with oneself that is associated with reasoning, self-regulation, and self-awareness. Moreover, incorporating overt inner speech in robotic systems can contribute to developing trust between humans and robots. When a robot's inner reasoning and evaluations are made audible to its human partner, it enables the human to understand the robot's decision-making process better, thus making the robot's behaviour more predictable and legible and fostering transparency.

To address transparency in human-robot interaction during learning, this study designs, evaluates and compares three naturalistic mechanisms for generating transparent robot behaviours. Specifically, we investigate the effects of verbalization of the inner speech in reaction to rewards, emotions, and a hybrid mechanism on robot behaviour during learning. The study examines the impact of these mechanisms on human perception of robots and their transparency levels.

II. Related Work

Reinforcement Learning (RL) techniques integrated into robots often appear as black-box learning systems to humans, making it difficult for users to understand the robot's decision-making process. Although recent HRI literature has begun addressing this issue, the number of studies remains limited and has only recently garnered attention.

Pynadath *et al.* [11] developed a decision-tree-based explainable model for RL to enhance transparency in the robot's decision-making. They conducted a study assessing the impact of text explanations generated using decision trees on fostering transparency and trust calibration. Although the results showed a positive impact on the robot's transparency, the study was conducted in simulated environments and did not evaluate the human perception of the robots.

Hirschmanner *et al.* [12] presented a study with a Pepper robot that learns object and action labels, and they investigated two extensions geared towards increasing the robot's transparency during learning. The first extension utilizes deictic gestures (i.e., pointing and gazing) to communicate knowledge, while the second extension shows the current state of the lexicon on the robot's tablet. In their study, they did not observe a significant increase in transparency, but users reported a higher perception of control and perceived learning success the more they interacted with the system.

Broekens *et al.* [13] proposed a computational model of emotions as a mapping between RL primitives and emotional labels. The authors showed that the proposed computational model allows us to compare the dynamics of RL primitives with the ones described in the psychological and behavioural literature on emotions. However, the work of Broekens et al. [13] mainly focuses on the possible roles of emotions for action selections via prediction and anticipation of future outcomes in a task involving a sequence of actions. The communicative role of emotion in robots that learn, using RL, in interaction with humans was also proposed in [14]. Starting from their computational model, the authors propose the use of such emotions to select the appropriate emotional expressions and to communicate to humans the state of their learning process. In that work, joy/distress signals are linked to a positive/negative temporal difference error, while emotions such as hope/fear are linked to anticipation mechanisms. However, no evaluation was provided on transparency.

Matarese *et al.* [7] proposed a model to enhance a robot's transparency during Interactive RL tasks by incorporating non-verbal emotional and behavioural cues into a humanoid robot. In their model, human feedback served as the RL algorithm's reward, and the robot exhibited emotional and behavioural responses based on its learning progress. Although the results indicated that people preferred interacting with an expressive robot over a mechanical one, the model led to misinterpretations when the robot expressed doubt or uncertainty, adversely impacting the robot's transparency.

In another work, Angelopoulos *et al.* [6] introduced two categories of a robot's emotional and behavioural reactions, one related to the robot's learning process and another in response to user feedback. Their research demonstrated increased transparency when the robot relied on both these categories. However, the authors focused exclusively on non-verbal cues and did not investigate the role of verbal cues in their implementation.

Considering these studies and their limitations, further exploration is still needed into developing and implementing natural and efficient verbal and non-verbal emotional behaviours/reactions in robots during the learning process to improve transparency and facilitate seamless HRI.

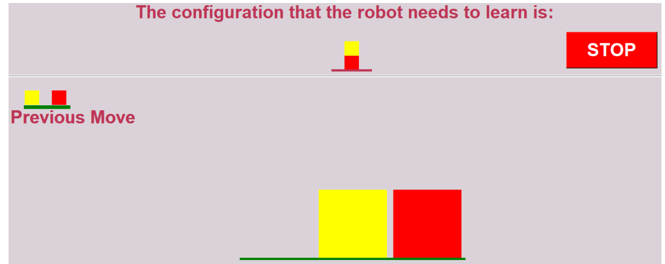


Fig. 1: The interface where the human can observe each step.

III. Methods

The temporal Difference (TD) technique, a widely utilized RL method, was employed to estimate the value of a state or action grounded in the expected reward and the expected value of the subsequent state or action. In RL, an agent learns by interacting with its environment, performing actions, and receiving rewards. The agent's goal is to maximize the cumulative reward over time. To accomplish this, the agent estimates state or action values using a value function, which the TD technique updates based on the difference between expected rewards and current value function estimates.

The TD error, which represents the discrepancy between the current state value estimate and the actual reward added to the discounted value estimate of the following state, is computed according to Equation 1. The TD error is also a natural measure of the divergence between the agent's expected and actual rewards.

$$TD_{error} \leftarrow R_{t+1} + \gamma \cdot V(S_{t+1}) - V(S_t) \tag{1}$$

Moreover, it is used to update the value estimate for the current state by adding the TD error multiplied by a learning rate α to the current value estimate, as shown in Equation 2.

$$V(S_t) \leftarrow V(S_t) + \alpha \cdot TD_{error} \tag{2}$$

One of the primary advantages of TD is that it can update the value function after every time step, making it more computationally efficient and allowing the agent to learn online. The TD error also provides a natural measure of the discrepancy between the agent's expected and actual rewards.

A. The Learning Task

We chose a modified version of the Towers of Hanoi problem as a learning task. The problem was modified to facilitate the robot's policy learning. In our version, the learning agent must move two cubes from a random initial position to a final one chosen by the human. The agent is allowed to move one cube per time step, and it can move each cube to the first, second, or third pillar. Furthermore, the learning agent receives a reward of 1 when all cubes are in the final configuration, a step cost of $-.001$ per time-step, and a negative reward of -1 if the agent has exceeded the maximum number of steps per episode. Participants oversaw the learning task, including the agent's previous move and the goal configuration, and provided their inputs through a graphical interface.

IV. The Proposed Approach

This work, as in our previous ones [6], [7], builds on the concept that an emotional reaction can be linked to feedback [15], neural temporal difference assessment and, therefore, to reward processing in RL and TD errors [13]. As explored in the literature, positive/negative emotional reactions are linked to positive/negative temporal difference errors. They are generated when the outcome of an action is better/worst than expected and would increase the expected value of that option. Moreover, affective reactions have an impact on the learning process also in terms of physiological arousal. Hence, the arousal of the emotional reaction is linked to the magnitude of such error with higher/lower arousal corresponding to bigger/smaller TD errors. Moreover, since, according to [16], emotion awareness obeys Weber's law of perception, the intensity of emotional reactions is linked to the TD error variation with a logarithmic association.

Here we are interested in using emotional and behavioural responses, as linked to the TD evaluation, to be used to make an agent's learning process more transparent to the observing user. Considering these factors, we aim to evaluate effective ways of displaying and signalling such emotional reactions. To create a more interactive experience for the users, three behavioural mechanisms were integrated into a physical Furhat robot, which was situated alongside the user interface as described in the following. This allowed the robot to provide real-time feedback and foster a more engaging and transparent learning environment for the observers.

*a) **Behavioural Mechanism 1 (BM1 - Inner Speech)**:* As the display of emotions on robots with non-verbal behaviour only may be difficult to be interpreted and could influence humans in unintended ways [7], here we consider the possibility of verbalization of an emotional or behavioural reaction. With respect to the use of inner speech to provide an explanation of the reasoning process [17], or in RL [11] to explain the learned policy, here we use inner speech for expressing verbal or para-verbal emotional feedback (i.e., "good", "great", "mhmm") with also linguistic acknowledgement of the action outcome with different intensity modulations (i.e., "I'm getting better", "I'm doing great"). In [11], the authors also included an acknowledgement of the action outcome before explaining its learning, but such acknowledgements were constant and did not vary during the learning progress. Our use of inner speech is associated with the robot's learning process, and therefore, it is elicited after each robot action.

The sentences used in the experimentation are reported in Table I. These sentences reflect different stages of the learning process, from acknowledging progress and maintaining motivation to expressing dissatisfaction and recognizing the need for strategy adjustments. Indeed, in HHI, various factors can trigger inner speech, such as emotional contexts, external objects, and the individual's internal state. Based on the specific trigger, distinct types of inner speech may manifest [10]. Therefore, alternative inner-talk sentences were provided for each condition within the same TD error value range, enabling the robot to switch between them during consecutive occurrences of identical TD error values. This approach mitigates monotony in the robot's behaviour and helps maintain observer engagement.

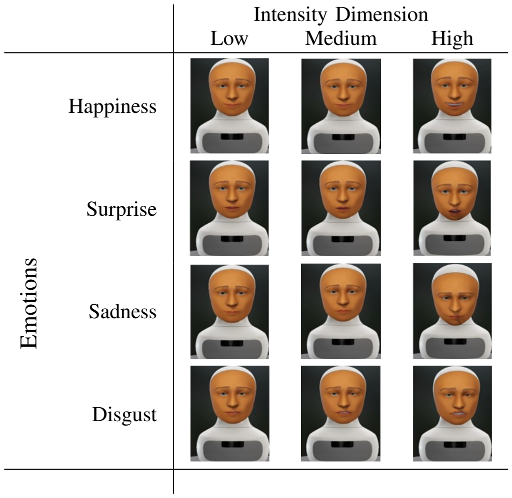
*b) **Behavioural Mechanism 2 (BM2 - Facial Emotions)**:* With respect to other social cues to display emotions, facial expressions are the ones better recognized in HHI e HRI. To have realistic facial expressions, in this work, we relied on the use of a Furhat robot. Here, facial expressions are associated with the robot's learning process and are elicited both during the learning process and after each robot action (see Table II). As affective behaviours, we considered negative emotions, such as disgust, sadness, and anger, and positive ones as surprise, happiness, confidence, and elation.

Disgust and sadness are used in the same TD error range to represent the agent's negative emotional responses to unfavourable learning experiences. With different simulated intensities, these emotions capture different aspects of dissatisfaction or disappointment when the agent receives a lower reward than expected or faces undesired outcomes. Disgust conveys the agent's aversion to its performance and may prompt the reconsideration of its current approach, while

TABLE I: robot's behaviour in terms of the Temporal Difference Error.

Temporal Difference Error	**Robot's Behavioural Mechanisms**		
	Inner Speech	**Facial Emotions**	**Hybrid Model**
$0 \leq TD \leq 10^{-9}$	-Not bad. -I am learning in baby steps.	Confidence	Surprise (intensity: low) Happiness (intensity: low)
$10^{-9} < TD \leq 10^{-5}$	-Good. -I am getting better.	Surprise (intensity: low) Happiness (intensity: low)	Surprise (intensity: medium) Happiness (intensity: medium)
$10^{-5} < TD \leq 10^{-1}$	-Very good. -I am learning.	Surprise (intensity: medium) Happiness (intensity: medium)	-Good. -I am learning in baby steps.
$10^{-1} < TD \leq 1$	-Excellent -I am doing great.	Surprise (intensity: high) Happiness (intensity: high)	-I am doing great. -Excellent.
$TD > 1$	-WOW, exploring is good for me.	Elation	Elation
$-9 \cdot 10^{-4} \leq TD < 0$	-Mhmm - I could have made a better choice.	Disgust (intensity: low) Sadness (intensity: low)	Disgust (intensity: low) Sadness (intensity: low)
$-10^{-3} \leq TD < -9 \cdot 10^{-4}$	-This is not good. -That is not good at all.	Disgust (intensity: medium) Sadness (intensity: medium)	Disgust (intensity: medium) Sadness (intensity: medium)
$-1 \leq TD < -10^{-3}$	-It was not a good move at all. -This move was even worse.	Disgust (intensity: high) Sadness (intensity: high)	-This move was even worse. -It was not a good move at all.
TD < -1	-From all these mistakes, I should have learned something.	Anger (intensity: high)	-This is not really good.

TABLE II: Exemplar of stimuli projected onto Furhat robot.



sadness signifies a sense of regret that can serve as a cue for having to adjust its strategy. By including both emotions in the same TD range, the agent can alternate between them, providing a more nuanced and varied response to negative learning experiences.

Happiness and surprise are also used in the same TD error value range to represent the agent's positive emotional responses to favourable learning experiences. These two emotions capture different aspects of satisfaction or excitement when receiving a higher reward than expected. Happiness conveys the agent's contentment with its performance and reinforces its learning progress, while surprise signifies unexpected or novel outcomes that might encourage further exploration. Moreover, while elation is an emotion, it is a strong affective signal acknowledging the agent's success in achieving a reward that greatly exceeds its expectations. This emotion represents a powerful positive reinforcement for the agent and encourages the exploration of new strategies or actions. It is used in the scenario to convey a sense of extreme happiness or excitement, aligning with the overall goal of creating a relatable and understandable learning process [18].

*c) **Behavioural Mechanism 3 (BM3 - Hybrid Model)**:* The Hybrid Model is a combination of inner speech and emotional expression during the learning process to create a more flexible and engaging learning experience for human observers. With the predefined TD error value ranges as depicted in Table I, the agent can alternate between emotions (for small and frequent variations) and inner speech expressions (for bigger improvements), offering a nuanced and diverse response to learning experiences. This adaptability enhances the human-like aspect of the robot's behaviour, prevents monotony, and promotes a more engaging and transparent learning process for observers. Using emotions for smaller TD error value ranges allows the agent to convey subtle changes in its performance, both positive and negative, in a non-verbal manner. They offer a quick and easily discernible way to communicate the agent's progress without overwhelming the observer with too much verbal information. On the other hand, inner speech is utilized for larger TD error value ranges to emphasize significant changes in performance, warranting explicit verbal acknowledgement.

V. Experiments

We investigated which behavioural mechanisms (BM1, BM2, or BM3) could convey a more transparent state of the robot during learning, and how the human observer perceives them. Based on existing literature, we hypothesized that Inner Speech (BM1) is a more transparent mechanism than Emotions (BM2) (**Hypothesis 1**). By listening to the robot's inner evaluation, its behaviour is anticipated to become more legible and less unpredictable, thus influencing transparency growth. Indeed, Geraci et al. [19] posited that inner speech could impact transparency levels. Finally, it is expected that the Hybrid Model (BM3) provides the most liked and transparent communication of the robot's state than Inner Speech (BM1) (**Hypothesis 2**). This expectation arises because the Hybrid Model combines multiple behavioural mechanisms, offering a complete understanding of the robot's state. We believe that excessive verbal communication, as in the case of Inner Speech (BM1), may lead to information overload and potential annoyance.

A. *Experimental settings and procedure*

The transparency of the proposed behaviour models was assessed through a between-subjects design. Participants were randomly assigned to different conditions. We used a video conferencing system to make participants remotely observe the behaviour of the robot and interact with the system. An experienced experimenter oversaw the human interaction with the robot to guarantee seamless execution.

Upon arrival, participants were provided with an informed consent form detailing the experiment's objectives and procedures. They were then introduced to the experimental environment and the physical robot. Participants were instructed to choose their preferred final configuration of cubes and indicate the experiment's completion by pressing the stop button (in the upper right corner of Figure 1) when they believed that the robot had learned the optimal path between the randomly selected initial configuration and the participant-selected final configuration.

B. *Measures*

HRI literature presents various interpretations of transparency [20]. Our study aims at using a broad understanding of transparency. The interpretation we adopt aligns closely with the characterization provided in our prior work [6]. Considering the aforementioned, we assessed transparency at each experimental session. The physical key metric used to evaluate transparency was the stop button. Specifically, participants were instructed to press this button once they had gained confidence that the robot had successfully learned the task at hand. This moment of confidence would typically coincide with a clear understanding of the robot's actions during the learning process, a comprehension of the reasons behind those actions, and an anticipation that the robot's future moves would be consistent with the desired task

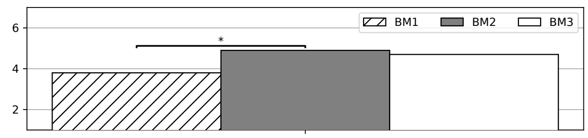


Fig. 2: Average number of epochs believed necessary for the robot to learn the optimal path.

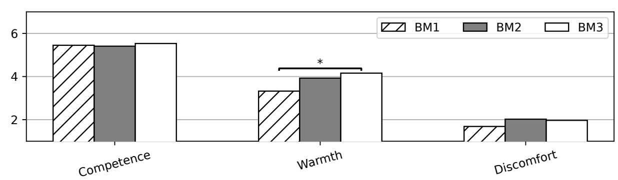


Fig. 3: RoSAS items, * indicates $p < 0.05$.

outcome. Finally, we also measured HRI safety, comfort, and reliability factors, as they have been identified as factors for transparency in prior studies [21].

Before the experiment, we collected participants' demographic information, including age, gender, education, and previous experience with robots. Participants also responded to a 5-point Likert scale question assessing their fear of machines becoming uncontrollable, which helped to evaluate potential negative biases towards robots. To examine the robot's behaviour and its impact on participants' perception of the social attributes of the robot, we administered a post-interaction survey at the end of each experimental session. This survey consisted of 7-point Likert and 18 cognitive-differences scale questions based on the Robot Social Attribute Scale (RoSAS) [22].

VI. Results

We recruited 108 participants, equally distributed across three conditions. The participants stated to be 72 males, 34 females, and two individuals preferred not to disclose their gender. The age range of participants was between 18 and 72 years old (28.5 ± 10.9), and none of them was familiar with the study's setup. The majority of participants (71.3%) had no prior experience interacting with robots, while 28.7% reported having previous experience. We evaluated participants' biases towards robots, and the scores indicated that participants exhibited no significant negative biases towards robots (2.4 ± 1.1); consequently, we did not exclude any participants. The participant pool provided an effect size of d=0.25 with .80 power at an alpha level of .05.

A. Reliability Analysis

A Cronbach's alpha test was performed to assess the internal reliability of the questionnaire. The test results indicated that the Warmth dimension of the RoSAS questionnaires had internal reliability of $\alpha_{Warmth} = 0.79$, the Competence dimension was found to be $\alpha_{Competence} = 0.84$, and the Discomfort dimension had internal reliability of $\alpha_{Discomfort} = 0.72$.

B. Results of Transparency

At first, we focused on evaluating the time at which participants pressed the stop button during learning. An independent samples t-test was conducted and revealed a statistical difference between BM1 and BM2 regarding the pressing of the stop key, $t(66.567) = -2.036, p = 0.046$. Figure 2 presents the average number of epochs participants believed necessary for the robot to learn the optimal path. No statistical differences were found in the average difference between the effective convergence of the Q-Table and the participants' pressing of the stop button. Our observations indicate that participants pressed the stop button as soon as the robot learned the optimal path. Specifically, the robot equipped with inner speech was perceived as more transparent than the other two behavioural mechanisms. This suggests that inner speech enabled participants to better understand when the robot had learned the optimal path.

Regarding the safety, comfort, and reliability factors, statistical differences were not observed ($p > 0.7$). We believe that the absence of statistical differences may be attributed to the novelty effect, as 71.3% of participants had no prior experience with robots leading to evaluating all the conditions positively [23].

C. Robotic Social Attributes Scale Ratings

An Independent Samples T-test was conducted with a 95% confidence level to investigate whether a statistical difference exists among the behaviour mechanisms. The results, as depicted in Fig. 3, indicate that there is a statistical difference in the average responses across the conditions in the "Warmth" scale, specifically between BM1 and BM3 ($t(68.200) = -2.258$, $p = 0.027$).

Further exploratory analyses were conducted to compare ratings for each of the RoSAS subitems between the conditions, as shown in Fig. 4. Statistical difference was found in the subitem of awkwardness, and participants perceived the robot as more awkward when utilizing BM2 (emotional behaviour) compared to BM1 (inner speech) ($t(54.736) = -2.141$, $p = 0.037$). The participants considered the robot with BM3 (hybrid behaviour) to be more capable of feelings and more emotional in comparison to BM1 ($t(67.011) = -2.065$, $p = 0.043$ and $t(66.851) = -2.691$, $p = 0.009$ respectively). However, no statistical differences were found in the remaining subitems of the RoSAS scale.

D. Evaluation of the Experimental Results

The aim of this study was to evaluate the transparency of robot learning. Our findings support **Hypothesis 1** that Inner Speech (BM1) is a more transparent mechanism than Emotions (BM2). Participants pressed the stop button as soon as the robot learned, indicating high transparency capabilities. Regarding **Hypothesis 2**, our results demonstrated higher ratings for the robot's social perception in the Hybrid Model (BM3), especially in the Warmth scale. The addition of emotional behaviour to inner speech made the robot perceived as more capable of feelings and more emotional.

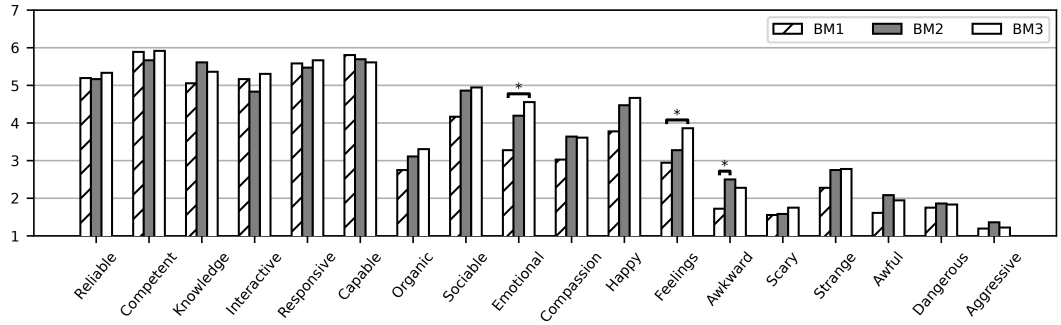


Fig. 4: RoSAS subitems, * indicates $p < 0.05$.

However, these differences were not enough to confirm Hypothesis 2 and additional evaluations are needed.

VII. Conclusions

In this study, we investigated the effects of various behavioural mechanisms on robot transparency. We focused on the use of inner speech, emotional behaviours, and a hybrid model which combined the previous two. Our experimental findings showed that the inner speech mechanism (BM1) significantly improved transparency compared to the emotional behaviour mechanism (BM2). Although the hybrid model (BM3) displayed higher ratings for anthropomorphism, particularly in the warmth dimension, it did not exhibit the expected improvement in transparency.

These results suggest that a robot endowed with inner speech can enhance transparency and improve the overall human-robot interaction experience. This aspect should be included in the design for robots' transparent behaviours to develop robotic systems that can communicate more effectively. Considering the satisfying results that this study produced, we plan to refine our model to consider other characteristics that may impact transparency, such as an adaptive selection of the presented mechanisms. Future research should also address the generalizability of our findings regarding anthropomorphic inferences with other robotic platforms and more complex learning scenarios.

References

[1] B. Nesset, D. A. Robb, J. Lopes, and H. Hastie, "Transparency in hri: Trust and decision making in the face of robot errors," in *Companion of the 2021 ACM/IEEE HRI*, 2021, pp. 313–317.

[2] D. Helbing, P. Molnár, I. J. Farkas, and K. Bolay, "Self-organizing pedestrian movement," *Environment and planning B: planning and design*, vol. 28, no. 3, pp. 361–383, 2001.

[3] A. Morin, "Inner speech," in *Oxford Companion to Consciousness*, Oxford University Press, 2009.

[4] G. Angelopoulos, A. Rossi, C. Di Napoli, and S. Rossi, "You are in my way: Non-verbal social cues for legible robot navigation behaviors," in *IEEE/RSJ IROS*, IEEE, 2022, pp. 657–662.

[5] G. Angelopoulos, F. Vigni, A. Rossi, G. Russo, M. Turco, and S. Rossi, "Familiar acoustic cues for legible service robots," in *31st IEEE RO-MAN*, IEEE, 2022, pp. 1187–1192.

[6] G. Angelopoulos, A. Rossi, G. L'Arco, and S. Rossi, "Transparent interactive reinforcement learning using emotional behaviours," *International Journal of Social Robotics*, 2023.

[7] M. Matarese, A. Sciutti, F. Rea, and S. Rossi, "Toward robots' behavioral transparency of temporal difference reinforcement learning with a human teacher," *IEEE Transactions on Human-Machine Systems*, vol. 51, no. 6, pp. 578–589, 2021.

[8] S. Asai, K. Yoshino, S. Shinagawa, S. Sakti, and S. Nakamura, "Eliciting cooperative persuasive dialogue by multimodal emotional robot," in *12th Intern Work on Spoken Dialogue System Technology, IWSDS 2021, Singapore*, Springer, 2022, pp. 143–158.

[9] S. Wallkötter, S. Tulli, G. Castellano, A. Paiva, and M. Chetouani, "Explainable embodied agents through social cues: A review," *J. Hum.-Robot Interact.*, vol. 10, no. 3, Jul. 2021.

[10] A. Pipitone and A. Chella, "What robots want? hearing the inner voice of a robot," *Iscience*, vol. 24, no. 4, p. 102 371, 2021.

[11] D. V. Pynadath, N. Gurney, and N. Wang, "Explainable reinforcement learning in human-robot teams: The impact of decision-tree explanations on transparency," in *31st IEEE RO-MAN*, IEEE, 2022, pp. 749–756.

[12] M. Hirschmanner, S. Gross, S. Zafari, B. Krenn, F. Neubarth, and M. Vincze, "Investigating transparency methods in a robot word-learning system and their effects on human teaching behaviors," in *30th IEEE RO-MAN*, IEEE, 2021, pp. 175–182.

[13] J. Broekens, E. Jacobs, and C. M. Jonker, "A reinforcement learning model of joy, distress, hope and fear," *Connection Science*, vol. 27, no. 3, pp. 215–233, 2015.

[14] J. Broekens and M. Chetouani, "Towards transparent robot learning through tdrl-based emotional expressions," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 352–362, 2021.

[15] R. F. Baumeister, K. D. Vohs, C. N. DeWall, and L. Zhang, "How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation," *Personality and Social Psychology Review*, vol. 11, no. 2, pp. 167–203, 2007.

[16] R. Berkovich and N. Meiran, "Pleasant emotional feelings follow one of the most basic psychophysical laws (weber's law) as most sensations do," *Emotion*, 2022.

[17] A. Pipitone, A. Geraci, A. D'Amico, V. Seidita, and A. Chella, "Robot's inner speech effects on trust and anthropomorphic cues in human-robot cooperation," *arXiv preprint arXiv:2109.09388*, 2021.

[18] M. K. Johnson, "Joy: A review of the literature and suggestions for future directions," *The Journal of Positive Psychology*, vol. 15, no. 1, pp. 5–24, 2020.

[19] A. Geraci, A. D'Amico, A. Pipitone, V. Seidita, and A. Chella, "Automation inner speech as an anthropomorphic feature affecting human trust: Current issues and future directions," *Frontiers in Robotics and AI*, vol. 8, p. 620 026, 2021.

[20] S. Y. Schött, R. M. Amin, and A. Butz, "A literature survey of how to convey transparency in co-located human–robot interaction," *Multimodal Technologies and Interaction*, vol. 7, no. 3, p. 25, 2023.

[21] C. Lichtenthäler, "Legibility of robot behavior," Ph.D. dissertation, Technische Universität München, 2014.

[22] C. M. Carpinella, A. B. Wyman, M. A. Perez, and S. J. Stroessner, "The robotic social attributes scale (rosas) development and validation," in *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction*, 2017, pp. 254–262.

[23] M. Koch, K. von Luck, J. Schwarzer, and S. Draheim, "The novelty effect in large display deployments–experiences and lessons-learned for evaluating prototypes," in *Proceedings of 16th European conference on computer-supported cooperative work-exploratory papers*, European Society for Socially Embedded Technologies (EUSSET), 2018.