Using Theory of Mind in Explanations for Fostering Transparency in Human-Robot Interaction*

Georgios Angelopoulos[1[0000−0001−9866−8719]], Pasquale Imparato[2], Alessandra Rossi[1,2[0000−0003−1362−8799]], and Silvia Rossi[1,2[0000−0002−3379−1756]]

[1] Interdepartmental Center for Advances in Robotic Surgery - ICAROS,
[2] Department of Electrical Engineering and Information Technologies - DIETI,
University of Naples Federico II, Naples, Italy
{georgios.angelopoulos,alessandra.rossi,silvia.rossi}@unina.it

Abstract. In human-robot interaction, addressing disparities in action perception is vital for fostering effective collaboration. Our study delves into the integration of explanatory mechanisms during robotic actions, focusing on aligning robot perspectives with the human's knowledge and beliefs. A comprehensive study involving 143 participants showed that providing explanations significantly enhances transparency compared to scenarios where no explanations are offered. However, intriguingly, lower transparency ratings were observed when these explanations considered participants' existing knowledge. This observation underscores the nuanced interplay between explanation mechanisms and human perception of transparency in the context of human-robot interaction. These preliminary findings contribute to emphasize the crucial role of explanations in enhancing transparency and highlight the need for further investigation to understand the multifaceted dynamics at play.

Keywords: Human-Robot Interaction · Explanations · Transparency

1 Introduction

Robots are engineered with specific tasks and objectives in mind; however, their actions may not always be readily understandable to humans. This lack of understanding can cause users to overestimate a robot's capabilities, a phenomenon known as overtrust [11]. Furthermore, the physical design elements of robots, encompassing their appearance and vocal attributes, significantly mould users' perceptions and expectations, highlighting the importance of considering Theory of Mind principles in designing robots that can better adapt to users' mental states.

* This work has been supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No955778 (G. Angelopoulos), by the Italian Ministry for Universities and Research (MUR) under the grant FAIR (MUR: PE0000013) (S. Rossi), and Italian PON R&I 2014-2020 - REACT-EU (CUP E65F21002920003) (A. Rossi).

To tackle these challenges, researchers have turned their attention to the use of verbal explanations to elucidate the decision-making processes of black-box algorithms. These techniques can be adapted to expound upon robotic actions, rendering them more comprehensible to human users [3]. The explanations proffered by robots hold substantial influence over how users perceive and engage with them since they can augment the anthropomorphic qualities of robots, thereby endowing robots with a more dynamic and human-like demeanour [12].

In this context, we believe that it is essential to consider the concept of the Theory of Mind with a robot providing explanations. Theory of Mind refers to the ability to attribute mental states, such as beliefs, intentions, and desires, to oneself and others. It has been shown that users might attribute mental states to a robot during the interaction [5]. On the contrary, the capability of reasoning on the user's possible mental states increases adaptability and efficiency [13]. Indeed, robots that can adapt their explanations based on the user's beliefs and knowledge levels can create a more intuitive and human-like interaction experience [8]. This adaptability enhances the robot's ability to communicate effectively and fosters a sense of understanding between the user and the robot. This understanding can lead to more contextually relevant explanations, which, in turn, could contribute to increased transparency.

This paper provides a contribution to exploring strategies to enhance Human-Robot Interaction (HRI) transparency and efficacy. We delve into the complex decision-making process that underpins selecting aspects of a robot's behaviour to be elucidated to users. Furthermore, we comprehensively examine the correlation between explanations that consider the human beliefs and the resultant perception of transparency. These preliminary findings provide insights into the ever-evolving landscape of HRI and towards advancing collaboration and transparency in robot behaviour.

2 Related Work

The importance of providing clear explanations for robot actions is widely acknowledged in HRI, yet a comprehensive understanding of how these explanations affect user perception remains a topic of ongoing investigation.

Ambsdorf *et al.* [1] conducted an online study to explore the impact of explainable robots on human perception. They designed a scenario where two simulated robots engaged in a competitive board game. One of the robots explained its moves, while the other simply announced them. The results revealed that the robot providing explanations was perceived as more dynamic and human-like. However, it also raised an important point that humans might still have reservations about trusting a robot's ability to perform tasks, even when it offers explanations.

Stange *et al.* [16] proposed an architecture for the social robot Pepper that allows it to interact autonomously with users and explain its behaviour based on the user's verbal requests during the interaction. Although their architecture

showed promise for creating robots with explainable autonomous behaviour, they did not investigate how these explanations influenced user perception.

Nikolaidis *et al.* [10] introduced a formalism that enables a robot to make decisions about whether to take action or provide explanations to a human teammate optimally. They employed verbal commands to guide humans and used state-explaining actions, where the robot explained its internal state while performing the task. Their study found that issuing verbal commands was the most effective way to communicate objectives while maintaining user trust in the robot.

In these studies, while the positive impact of explanations on HRI is evident, it is crucial to note that understanding human knowledge and beliefs and tailoring explanations accordingly remains a challenge to be explored. This aspect of considering the user's existing knowledge during explanations can significantly influence how humans perceive and trust robots in various tasks and scenarios.

3 Proposed Approach

To enhance transparency and mutual understanding within HRI, our proposed approach draws inspiration from the recent work by Sreedharan *et al.* [15], and centres on the concept of generating explanations as a process of reconciliation between plans generated from different world models. This approach is particularly relevant in scenarios where the interaction involves dynamic elements. In this dynamic scenario, we employ a framework rooted in knowledge representation and automated planning. Specifically, we leverage the ROSPlan framework, a robust tool for knowledge-based planning [4] and utilize Planning Domain Definition Language to model the knowledge states of both the robot and the human within the interaction context.

In this context, we consider a dynamic scenario where a change occurs within the environment during the task. Importantly, the robot's knowledge is updated to reflect this new information, whereas the human remains unaware of this change. This shift creates a fundamental disparity in their knowledge states and perspectives. In response to these dynamic changes, our approach employs the Fast Forward algorithm to derive optimal sequences of actions for both entities. The robot plans its actions while considering the updated information, whereas the human operates with their previous knowledge. These differing knowledge states become pivotal in shaping the explanations provided by the robot.

The robot's explanations are adapted to accommodate the cognitive divergence resulting from the changing environment. When the robot explains its actions, it considers both its own and the human's knowledge. For instance, if the robot reaches the goal of a task, it offers an explanation that aligns with the human's perspective. Considering their differing knowledge states, this could ensure that the robot's actions are more transparent and comprehensible to humans. Our approach relies in comparing these computed action sequences and recognizing instances where their knowledge diverges due to the changing of the environment. These disparities serve as valuable cues for tailoring explanations

that resonate with the human cognitive model, taking into account the evolving perspective caused by this change. Consequently, our approach facilitates the delivery of more constructive and contextually relevant explanations, considering the nuanced differences in knowledge and perception induced by dynamic elements. The combination of knowledge representation, automated planning, and a focus on cognitive divergence empowers the robot to elucidate its actions that align with the human's evolving perspective in these dynamic scenarios.

3.1 The Scenario

We employed a "fetch-and-carry" task [7], which is a simple interactive scenario to assess the explanations and the transparency derived from them in an HRI context. The robot navigated in the environment (see Figure 1), retrieved an item with a dynamic position (i.e., a drink), and delivered it to the virtual human.

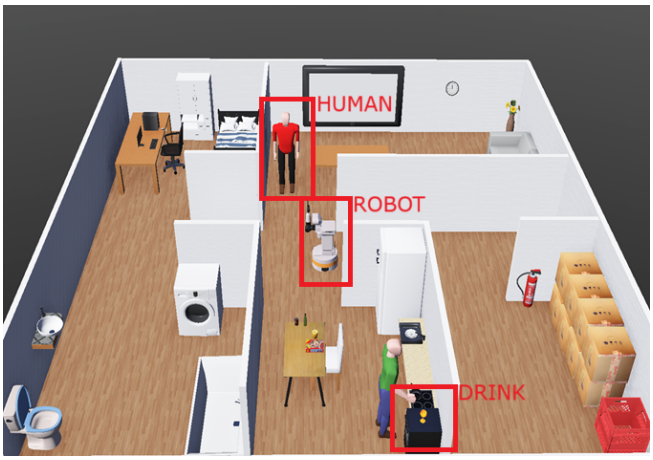

Fig. 1: The Employed Virtual Environment consists of five rooms.

To investigate the influence of considering human knowledge on the explanations in Transparency, we implemented three conditions:

- ***Condition 1*** **(C1)**: This condition served as the baseline and involved the robot not providing any explanations.
- ***Condition 2*** **(C2)**: In this condition, the robot offered an initial verbalization of its actions ("I have to get the drink") and a subsequent explanation after completing the actions ("The drink was in the living room").
- ***Condition 3*** **(C3)**: This condition integrated both the robot's initial explanation and the explanation after completing the action, along with knowledge about the human's preferences ("I have to take the drink, but it is not in the kitchen" and "The drink was not in the kitchen but was in the living room").

In light of the existing literature, we formulated the following hypotheses to guide our investigation:

- ***Hypothesis 1*** **(H1)**: The condition where the robot provides explanations (C2, C3) produces a more transparent mechanism than in the condition where a robot does not provide explanations. This hypothesis aligns with previous findings by Felzmann *et al.* [6], which demonstrated that providing explanations could enhance the transparency of robotic systems.
- ***Hypothesis 2*** **(H2)**: The condition in which the robot considers human knowledge into the explanations (C3) results in behaviour that is more transparent than the other mechanisms that do not consider human knowledge (C1, C2). This hypothesis builds on the findings of Milliez *et al.* [9], which suggested that adapting the explanations to the human's knowledge can lead to the robot being perceived as smarter.

4 User Study

An online between-subject study was conducted to assess the designed experimental mechanisms. We advertised the study via relevant email lists and on several social media platforms. We also used snowball sampling by asking participants to share the study information with interested friends and colleagues.

4.1 Procedure and Measurements

Initially, participants were required to complete a series of questions concerning their demographic information and prior experience with robots. Upon completing these questions, participants proceed to watch a video. The video displayed a static map labelled with room locations, the user's position, and the drink's location (see Figure 1).

Following this static map presentation, a second video segment commenced, wherein participants were presented with a first-person perspective from the human's point of view. This perspective allowed participants to have a partial observation of both the robotic entity and the surrounding kitchen environment (see Figure 2). During this video segment, the robot initiated the "fetch-and-carry" task. In the context of our experimental procedure, a deliberate interruption of the video was introduced as an essential component to gauge the transparency of the robot's actions during the task execution. This interruption was incorporated to solicit participants' evaluations of the robot's behaviour and happened when the robot exited their field of view of the human (in conditions C2 and C3, the robot had provided an explanation before the action).

To expound further on this approach, participants were prompted to provide assessments regarding the robot's Legibility, Predictability and Expectability, as these factors collectively contribute to the construct of Transparency [2]. More specifically, participants were asked to answer the following:

Fig. 2: The perspective of the virtual human.

- To what extent do you know why the robot moves the way it does? (*Legibility*).
- How well do you know what the robot will do next? (*Predictability*).
- To what extent does the robot behave as expected? (*Expectability*).

After responding, participants continued watching the remaining video. Upon completion, they were asked to fill out the Human-Robot Interaction Evaluation Scale (HRIES) Questionnaire assessing their perception of the robot [14]. But also 5-Likert scale questions regarding the overall Transparency of the robot:

- To what extent do you know why the robot moved the way it did? (*Legibility*).
- To what extent did you know what the robot would do next? (*Predictability*).
- To what extent did the robot behave as you expected? (*Expectability*).

It is important to notice that the video's content remained consistent across all conditions, with only the verbal explanations differing.

5 Results

In the conducted online study, we initially recruited a total of 178 participants. After careful screening, one participant was excluded due to being below the age of 18, while an additional seven participants were removed. These seven participants were disqualified based on the following criteria: duplicate survey submissions and the presence of extreme outlier scores across multiple measured variables. Consequently, the final set for analysis comprised 143 participants, consisting of 86 males and 57 females, with no non-binary or other genders. This resulted in an effect size of d = 0.25 with a power of 0.90 at an alpha level of 0.05. The participants exhibited a diverse age range spanning from 18 to 60 years (Mean=39.16, Std. Deviation=15.84). The majority of respondents (61.5%) reported no prior experience with robots. Furthermore, we observed that they had no negative bias towards robots (Mean=2.44, Std. Deviation=1.05).

5.1 System's Transparency

A series of t-tests were conducted to examine the differences in the characteristics of Transparency between the different conditions. Within the context of C1 and C2, statistically significant variations were observed in the Legibility of the robot's actions before pausing ($t(76.074) = -4.979, p < .001$), as well as in the Predictability ($t(94) = -7.143, p < 0.001$) and Expectability ($t(94) = -3.662, p < .001$) of its behaviour prior to a pause. Similarly, post-pausing, significant differences were detected in Legibility ($t(80.731) = -4.252, p < .001$) and Predictability ($t(94) = -4.212, p < .001$), as well as Expectability ($t(94) = -3.455, p < .001$).

In the case of C1 and C3, noteworthy disparities were identified in the robot's Legibility before pausing ($t(78.552) = -4.776, p < .001$), Predictability before pausing ($t(97) = -5.173, p < .001$), and Expectability before pausing ($t(97) = -4.466, p < .001$). Post-pause, significant distinctions were also evident in Legibility ($t(80.053) = -3.332, p < .001$) and Expectability ($t(97) = -3.880, p < .001$), whereas Predictability exhibited a significant difference ($t(97) = -2.500, p = .014$).

Conversely, in the comparison of C2 and C3, no statistically significant differences were discerned in the Legibility before pausing ($t(89) = 0.052, p = .479$), Predictability before pausing ($t(89) = 1.589, p = .058$), or Expectability before pausing ($t(89) = -0.832, p = .204$). Nevertheless, post to a pause, no significant differences were found in Legibility ($t(89) = -0.557, p = .290$) or Expectability ($t(89) = -0.366, p = .358$). Post-pause Predictability exhibited a statistically significant difference ($t(89) = 1.786, p = .039$), indicating distinguishable outcomes between C2 and C3. For visual clarity, Figure 3 provides graphical representations of the data, illustrating the variations in Legibility, Predictability, and Expectability across the different experimental conditions.

In our analysis of Transparency, we sought to evaluate its degree both before and after the pause. The results are depicted in Figure 4. To assess the degree of Transparency before the pause, we employed a factor analysis approach that integrates three key components: Legibility, Predictability, and Expectability. We first evaluated the dataset's suitability for factor analysis by examining the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy, which yielded a KMO value of 0.644, indicating its appropriateness for this analytical technique. Subsequently, we examined the factor loadings derived from our comprehensive analysis. Specifically, our data unveiled factor loadings of 0.858 for Legibility, 0.715 for Predictability, and 0.860 for Expectability. These loadings underscore the significant contributions of all three constituent elements to the overarching construct of Transparency, with Expectability emerging as the most influential factor, closely followed by Legibility and Predictability.

To calculate the relative importance of each component, we normalize the standardized factor loadings by dividing them by the sum of all three standardized loadings. The resulting relative importance scores are as follows: $\mathbf{R}(\mathbf{L}) : 0.352$, $\mathbf{R}(\mathbf{P}) : 0.294$, and $\mathbf{R}(\mathbf{E}) : 0.354$. Utilizing these weights in a weighted sum formula, we obtain the pre-transparency score:

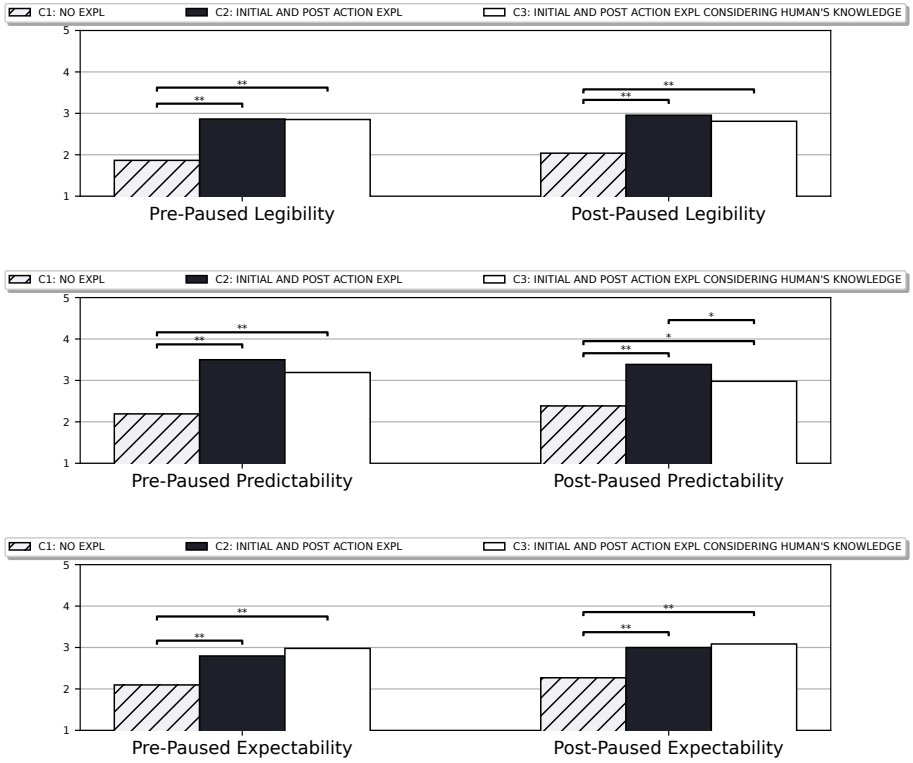

Fig. 3: Legibility, Predictability, and Expectability for each Condition (* for $p < 0.05$ and ** for $p \leq 0.001$).

$$T_{pre} = 0.352 \times \text{Legib}_{pre} + 0.294 \times \text{Predict}_{pre} + 0.354 \times \text{Expect}_{pre} \quad (1)$$

We conducted a similar factor analysis on the Transparency assessment after the pause. However, it is important to note that the KMO measure of sampling adequacy in this analysis yielded a value of 0.5, which is slightly below the preferred threshold for robust factor analysis. Nevertheless, the chi-square statistic of 29.670, coupled with a significance level less than 0.0001, indicated a statistically significant relationship among the variables, justifying the continuation of the factor analysis. From this, we obtained factor loadings that revealed the relative contributions of each variable to Transparency after the pause. Legibility had the most substantial influence with a factor loading of 0.901, followed by Expectability (0.813) and Predictability (0.523).

To calculate the relative importance of these components, we again normalized the standardized factor loadings and obtained the following relative importance scores: $\mathbf{R}(\mathbf{L}) : 0.402$, $\mathbf{R}(\mathbf{P}) : 0.234$, and $\mathbf{R}(\mathbf{E}) : 0.363$. Utilizing these weights in a weighted sum formula, we obtain the post-transparency score:

$$T_{post} = 0.402 \times \text{Legib}_{post} + 0.234 \times \text{Predict}_{post} + 0.363 \times \text{Expect}_{post} \quad (2)$$

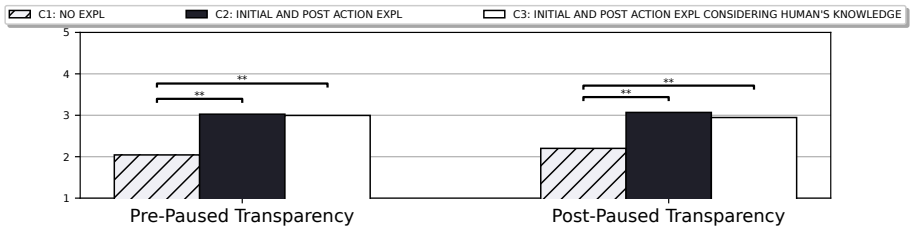

Fig. 4: Transparency for each Condition (* for $p < 0.05$ and ** for $p \leq 0.001$).

5.2 HRIES Ratings

In the initial phase of the investigation, an assessment of the internal reliability of the HRIES questionnaire was conducted. The results revealed Cronbach's alpha coefficients for the Sociability, Animacy, Agency, and Disturbance factors as follows: $\alpha_{Sociability} = 0.70$, $\alpha_{Animacy} = 0.75$, $\alpha_{Agency} = 0.71$, and $\alpha_{Disturbance} = 0.66$. However, it is noteworthy that the Cronbach's alpha coefficient for the Disturbance factor fell within an unacceptable range. In light of this finding, the item "Uncanny" was eliminated from the Disturbance factor. Subsequent to this adjustment, a revised Cronbach's alpha coefficient for the Disturbance factor yielded $\alpha_{Disturbance} = 0.81$.

To assess variations in HRIES factors across different conditions, a series of t-tests were conducted. The results are depicted in Figure 5. Notably, in the context of the *Sociability* dimension, a statistically significant difference emerged between C1 and C2 ($t(80.936) = -1.872, p = .032$), indicating that participants assigned more positive evaluations to C2 in terms of sociability. Conversely, in the *Animacy* dimension, no statistically significant differences were detected among the conditions, signifying consistent participant evaluations of animacy across conditions. In contrast, for the *Agency* factor, a statistically significant difference was noted between C1 and C2 ($t(94) = -1.810, p = .037$), with C2 receiving higher ratings in terms of agency. Lastly, in the *Disturbance* dimension, a statistically significant difference was observed between C1 and C2 ($t(81.790) = -2.147, p = .017$), indicating that C2 was associated with a higher level of perceived disturbance compared to C1. Sociability, Agency, and Animacy dimensions positively correlate with anthropomorphism [14]; the results thus suggest an increase in the robot's anthropomorphism in C2.

5.3 Evaluation of the Experimental Results

This study aimed to assess the mechanisms underlying explanations and their effect on transparency. Our results have confirmed Hypothesis 1 by showing that the provision of explanations by the robot (C2, C3) yields a more transparent mechanism than a robot without any explanatory discourse. Of particular significance is the observation that C3, where explanations were provided while taking into account participants' knowledge, received lower transparency ratings compared to C2, where explanations were given without considering participants'

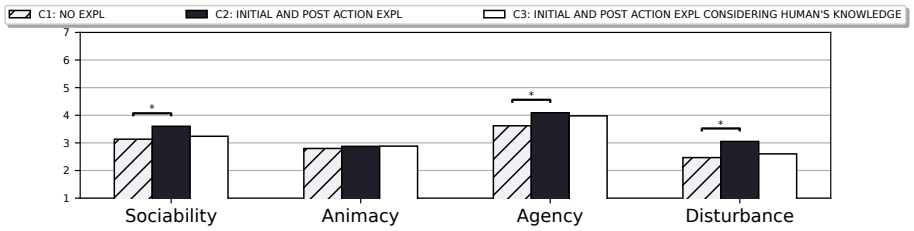

Fig. 5: Results of the HRIES questionnaire for each Condition (* for $p < 0.05$ and ** for $p \leq 0.001$).

knowledge. This outcome finds support in the higher rating received by C2 on the HRIES questionnaire, which is consistent with previous research [17] that emphasized the influence of transparency on anthropomorphism. Consequently, our findings did not confirm Hypothesis 2.

This divergence in transparency ratings between C2 and C3 may be attributed to several factors. Firstly, it is plausible that C3 introduced additional information, potentially resulting in a heightened cognitive load for participants. This increased cognitive burden could have rendered it more challenging for participants to effectively process and integrate the supplementary information, subsequently diminishing the perceived transparency of the explanations provided. Moreover, the combination of both the robot's and human's knowledge in C3 might have been perceived as intricate or redundant, thereby diminishing the clarity of the robot's intentions. Further investigation is needed to gain a more comprehensive understanding of these findings. These observations underscore the intricate interplay of various factors that influence the provision of explanations and the subsequent perception of transparency.

6 Conclusions

The work presented in this paper aimed at integrating explanatory mechanisms into human-robot interactions to enhance transparency and mutual understanding. Our study provided valuable insights into the complex dynamics at play when robots offer explanations, particularly in dynamic scenarios where the robot's knowledge differs from that of the human. We found that providing explanations significantly improves transparency compared to scenarios with no explanations. However, it was intriguing to note that considering participants' existing knowledge when crafting explanations did not necessarily lead to higher transparency ratings. These findings emphasize the need for a nuanced approach in designing explanations for robots and highlight the intricate balance between providing information and cognitive load.

References

1. Ambsdorf, J., Munir, A., Wei, Y., Degkwitz, K., Harms, H.M., Stannek, S., Ahrens, K., Becker, D., Strahl, E., Weber, T., et al.: Explain yourself! effects of explanations

in human-robot interaction. In: 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). pp. 393–400. IEEE (2022)
2. Angelopoulos, G., Di Martino, C., Rossi, A., Rossi, S.: Unveiling the learning curve: Enhancing transparency in robot's learning with inner speech and emotions. In: IEEE International Conf. on Robot and Human Interactive Communication (2023)
3. Angelopoulos, G., Rossi, A., L'Arco, G., Rossi, S.: Transparent interactive reinforcement learning using emotional behaviours. In: International Conference on Social Robotics. pp. 300–311. Springer (2022)
4. Cashmore, M., Fox, M., Long, D., Magazzeni, D., Ridder, B., Carrera, A., Palomeras, N., Hurtos, N., Carreras, M.: Rosplan: Planning in the robot operating system. In: Proceedings of the international conference on automated planning and scheduling. vol. 25, pp. 333–341 (2015)
5. Cucciniello, I., Sangiovanni, S., Maggi, G., Rossi, S.: Mind perception in hri: Exploring users'attribution of mental and emotional states to robots with different behavioural styles. International Journal of Social Robotics **15**(5), 867–877 (2023)
6. Felzmann, H., Fosch-Villaronga, E., Lutz, C., Tamo-Larrieux, A.: Robots and transparency: The multiple dimensions of transparency in the context of robot technologies. IEEE Robotics & Automation Magazine **26**(2), 71–78 (2019)
7. Kraus, M., Wagner, N., Untereiner, N., Minker, W.: Including social expectations for trustworthy proactive human-robot dialogue. In: Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization. pp. 23–33 (2022)
8. McKenna, P.E., Romeo, M., Pimentel, J., Diab, M., Moujahid, M., Hastie, H., Demiris, Y.: Theory of mind and trust in human-robot navigation. In: Proc. of the 1st Intern. Symposium on Trustworthy Autonomous Systems. pp. 1–5 (2023)
9. Milliez, G., Lallement, R., Fiore, M., Alami, R.: Using human knowledge awareness to adapt collaborative plan generation, explanation and monitoring. In: ACM/IEEE International Conference on HRI. pp. 43–50 (2016)
10. Nikolaidis, S., Kwon, M., Forlizzi, J., Srinivasa, S.: Planning with verbal communication for human-robot collaboration. ACM Transactions on Human-Robot Interaction (THRI) **7**(3), 1–21 (2018)
11. Rossi, A., Koay, K.L., Haring, K.S.: To err is robotic: Understanding, preventing, and resolving robots' failures in hri
12. Schött, S.Y., Amin, R.M., Butz, A.: A literature survey of how to convey transparency in co-located human–robot interaction. Multimodal Technologies and Interaction **7**(3), 25 (2023)
13. Shvo, M., Hari, R., O'Reilly, Z., Abolore, S., Wang, S.Y.N., McIlraith, S.A.: Proactive robotic assistance via theory of mind. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 9148–9155 (2022)
14. Spatola, N., Kühnlenz, B., Cheng, G.: Perception and evaluation in human–robot interaction: The human–robot interaction evaluation scale (hries)—a multicomponent approach of anthropomorphism. International Journal of Social Robotics **13**(7) (2021)
15. Sreedharan, S., Chakraborti, T., Kambhampati, S.: Foundations of explanations as model reconciliation. Artificial Intelligence **301**, 103558 (2021)
16. Stange, S., Hassan, T., Schröder, F., Konkol, J., Kopp, S.: Self-explaining social robots: An explainable behavior generation architecture for human-robot interaction. Frontiers in Artificial Intelligence **5**, 87 (2022)
17. Straten, C.L.v., Peter, J., Kühne, R., Barco, A.: Transparency about a robot's lack of human psychological capacities: effects on child-robot perception and relationship formation. ACM Transactions on Human-Robot Interaction (THRI) **9**(2), 1–22 (2020)