Humans' Spatial Perspective-Taking When Interacting with a Robotic Arm

Mouad Abrini, Malika Auvray, Mohamed Chetouani

Humans' Spatial Perspective-Taking When Interacting with a Robotic Arm

Mouad Abrini[1], Malika Auvray[1] and Mohamed Chetouani[1]

***Abstract*— Perceiving the environment from another person's perspective, in other words, being in someone else's shoes spatially, is not always an easy task. Perspective-taking can be even more challenging when working with a robot as a collaborator. The study reported here aims at investigating humans' level 2 spatial perspective-taking performance when interacting with a collaborative robotic arm through a novel in-person experiment. First, a robotic arm drew ambiguous shapes on a whiteboard and participants had to answer questions that require performing spatial perspective-taking. A metric was used to compute a score based on their responses. Second, participants completed the PTSOT, a test measuring spatial orientation and perspective-taking ability. The results revealed a correlation between the scores computed using our metric and those obtained in the PTSOT. This suggests the efficiency of our new setup and associated evaluation metric in assessing spatial perspective-taking skills in a human-robot interaction context, as well as the validity of our findings, in line with prior studies on perspective-taking.**

I. INTRODUCTION

Perspective-taking is a multidimensional construct referring to the ability to perceive a situation from someone else's point of view and to understand how our own actions and behaviors may be perceived by others [1]. It is an essential skill, that is used, either consciously or unconsciously, by individuals in their daily interactions [2]. Perspective-taking is often characterized along three dimensions: affective, cognitive, and spatial. Affective perspective-taking refers to the ability to understand the emotions and feelings of others. It is closely related to empathy and allows individuals to respond appropriately to the emotional states of other individuals. Cognitive perspective-taking refers to the ability to understand and interpret others' thoughts, behaviors and beliefs, and it can be referred to as theory of mind. Spatial perspective-taking corresponds to the ability to imagine how an object or a scene would appear from a perspective different from one's current physical viewpoint [3]. In this article, we focus on spatial perspective-taking in a human-robot interaction scenario. The field of Robotics is rapidly evolving, and robots are often introduced in environments where they have to interact with humans. This is usually the case of collaborative robots. One of the necessary components for a successful interaction with robots is spatial perspective-taking. This component has already been tackled by pioneering research [4] which showed that humans prefer the robot to take their perspective and act accordingly in case of spatial ambiguity. This brings an important question: What do we mean by spatial ambiguity? In the scenario of [4], the participant could only see one of the two objects, while the robot could see the two of them. If the participant asks the robot to go towards the visible object, this creates a spatial ambiguity. Our study focuses on shape spatial ambiguity and in particular on 2D drawn shapes that can be perceived differently depending on the observer's point of view. This task is coined as ambiguous grapheme perception, such as 6/9, b, d, p, q, whose perception varies as a function of the perspective that is adopted on the stimulus. To date, there are no studies evaluating human's spatial perspective-taking abilities when interacting on-line with a robot. It is of utmost importance to develop such methods, since robots need to be able to adjust their perception and movements based on humans' abilities. In the study reported here, a new task is proposed through which we can evaluate level 2 spatial perspective-taking abilities in a human-robot interaction context. In this task, a robotic arm draws an ambiguous 2D shape on a whiteboard. Then, the participants are required to answer questions on a tablet while performing spatial perspective-taking. The main aim of this study is to investigate the validity of this experimental method through which we can evaluate level 2 spatial perspective-taking in a human-robot interaction context. This method might prove useful to personalize human-robot interactions based on individual abilities.

* This work was not supported by any organization

[1]All authors are with Sorbonne University, CNRS, Institut des Systèmes Intelligents et de Robotique, Paris, France {name.surname}@sorbonne-universite.fr

II. RELATED WORK

A. Humans' spatial perspective-taking

According to previous findings in psychology, there are two levels of spatial perspective-taking [5], [6]. Level 1 spatial perspective-taking refers to the ability to understand that an object may be visible from a specific perspective but not from another one. This is usually referred to as a "I know what you see" situation [7]. This ability is usually explored through experiments where two individuals engage in interaction within a setting where certain objects are visible to one person and not to the other. For example, previous studies used a dot task where the participant has to count the number of dots that an avatar facing a wall can see or not [8], [9].

[1] and [10] have shown that the underlying mechanism of level 1 perspective-taking in humans works by imagining a line-of-sight between the object and the individual. The object is assumed to be occluded if the line-of-sight is interrupted by an interposed obstacle. It appears that children are able to develop level 1 perspective-taking abilities by

the time they reach the age of 2 [11].

Level 2 spatial perspective-taking on the other hand corresponds to the ability to understand and imagine a physical scene from a different viewpoint. This level of perspective-taking requires one to perform a virtual embodied rotation in order to move themselves to another position or viewpoint within the same environment and imagine what they would see from that new perspective. This would be a "I can see the world through your eyes" situation [7]. This level of perspective-taking can be investigated through multiple experimental setups. [12] for example, study level 2 spatial perspective-taking by asking participants how a number appears from the perspective of a human avatar. Level 2 spatial perspective-taking is considered to be more advanced than level 1. In fact, it seems to be developed between the years of 4 and 5 [13]. In addition, level 2 perspective-taking abilities has been shown to be dependent on sensory factors such as visual [14] and proprioceptive [15] deficits and on individual factors such as social intelligence and attachment style [16].

B. Spatial perspective-taking during human-robot interactions

Level 1 spatial perspective-taking was investigated by [17] in the context of human-robot interaction. In order to handle ambiguous situations, such as when a human requests a tool from a robot and the human can only see one tool while the robot can see two tools (including the one that is occluded from the human's perspective), the authors used a framework referred to as Polyscheme [18], [19], which helps resolve such scenarios. Polyscheme is a cognitive architecture used to model how humans use multiple methods of representation, reasoning and problem-solving. For example, in the context of [17], when two cones are available, the robot chooses the one visible by both agents respecting the principles of the least effort and joint salience [20].
Level 1 spatial perspective-taking was also investigated by [21] and [22] in a human-robot interaction context. The authors used a line-of-sight tracing (as presented in II-A) approach in order to infer if an object is visible or not. The only difference between [21] and [22] is that the former uses motion capture for the objects and human perception, and the latter presented a method that does not require the environment to be known.

Level 2 spatial perspective-taking was previously investigated by [23], in a human-robot interaction context where the humanoid robots used were the NAO robot and the Baxter. The robots had to reach or look at an ambiguous shape (6/9). The results revealed that humans adopted the robot's perspective more frequently in the reaching scenario than in the looking condition. In addition, robot's appearance influenced participants' performance, and perspective-taking abilities increased when the robot had a human-like appearance.

Some research works aims at granting robots with the ability to adopt human agents' perspective. This is done by attempts to equip robots with spatial perspective-taking ability. For example, in [24] referring expressions are generated while human agent's perspective is taken into account. Humans moved during the interaction, which changed their perspective on the objects. To allow updates on the perspective, the authors created a perspective-taking module that relies on users' coordinates (center of gravity) and viewpoints collected beforehand. The latter was recorded in the form of an image and coordinates in space. In [24], 8 viewpoints were collected. Then, to find the closest viewpoint to the participant, the authors calculated the Euclidean distance between the coordinates of the human (center of gravity) and the coordinates of the stored perspectives. The closest viewpoint is the one that minimizes that distance. The center of gravity was obtained using a pre-trained Mask-RCNN model [25].
A more robust method was developed by [22] that works under unconstrained and markless environments. First, the environment was mapped using Real-Time Appearance-Based Mapping (RTAB-MAP) [26]. Real-time object recognition was performed using deep learning algorithms. Then, to estimate what would the world look like to the human agent, the 3D point cloud map was transformed to the frame of the user that is estimated using head pose and gaze estimation algorithms.

III. Experiment setup and methods

A. Participants

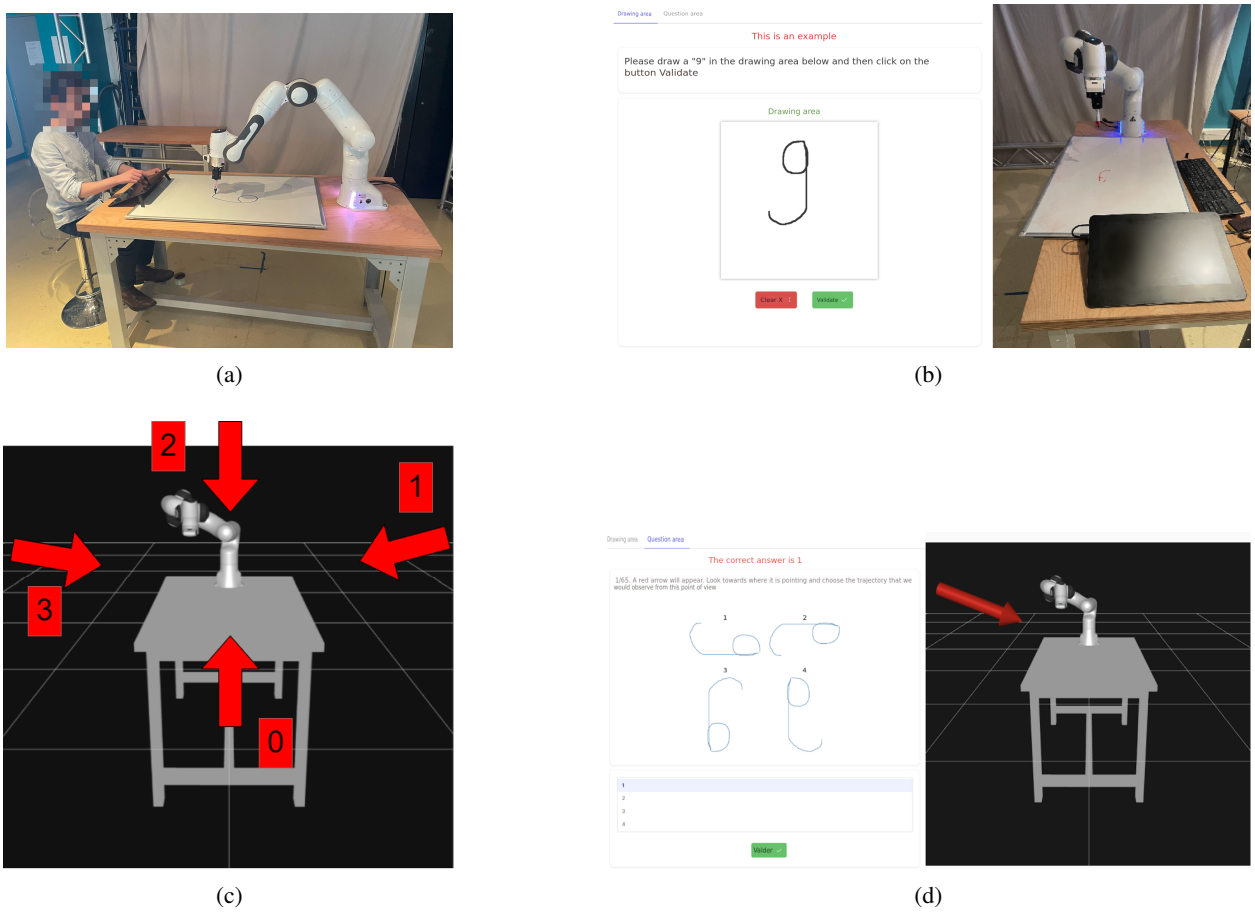
39 participants were recruited via the local information relay on cognitive sciences, which is a volunteer platform for experiments. The participants received 10 euros as compensation for the experiment that took on average 20 minutes to complete. The participants' ages ranged from 19 to 42 years (M = 29.3, SD = 6.6; 24 females, 15 males).

Ethics approval for the study was obtained from Sorbonne University Ethics Committee (Protocol CER-2022-068). Participants were informed about the experiment and were asked to provide their consent by signing a form before the study started. The experiment was performed in accordance with the ethical standards laid down in the Declaration of Helsinki (1991). Data collection was conducted anonymously. The participants were involved in two blocks of the experiment, as presented in sections III-B and III-C.

B. Block 1 of the experiment

Figure 1a shows the configuration that was adopted in this block. The participants sat in front of the robotic arm (Franka Emika Panda), outside the robot's workspace, in order to ensure safety. Hence, it is impossible for the robot to reach the participants. The participant interacts with the robot through the Wacom Cintiq pro 16 tablet.

The participants were asked to draw one of the three ambiguous shapes: "p","q" or "6". The robot then randomly reproduced the shape, either from its perspective or from the participants' one. These trials were repeated 60 times with



(a) (b)

(c) (d)

Fig. 1: *The configuration of the setup - (a): Participant sitting in front of Franka Emika robot, (b): The drawing zone as a subcomponent of the user interface, and the result of the drawing performed by the robotic arm, (c): Representation of the virtual positions defined in table I, (d): The user interface representing the multiple-choice question and the red arrow indicating the decentred left direction*

the different ambiguous trajectories to ensure consistency of results. In addition, the participants' response times (RTs) were measured. RTs provide an additional performance indicator, as longer time taken to provide a correct answer might indicate that the participants are less comfortable with this condition.

1) User interface: The interaction between the participants and the robotic arm was done via a Web Interface. This interface contains an area for drawing trajectories, a 3D visualization of the robot in real time developed with ros3djs [1] (Figures 1c and 1d), as well as the area in which the user will be asked to respond. The web interface represents the front-end of the application. A Web server was also developed. It manages the logic of the Web application and the communication between the front-end and the robot. The interface was developed using the Javascript framework VueJs. Figures 1b and 1d show a preview of the two main screens of the interface. The source code is available on Github[2].

In the example shown in figure 1b, the participant is asked to draw a "9". The result of the drawing by the robot is shown on the same figure. In fact, once the trajectory is drawn by the participant, the robot reproduces it on the whiteboard, either from its own egocentric perspective or from the perspective of the user. For the next step, the screen shown in figure 1d is presented to the participant.

A red arrow appears in the 3D visualization area. The participant is then asked to choose the trajectory that they would see if they were standing in the same direction as the arrow. The participant is not allowed to move away from their position. This means that they cannot stand up or move around the table to complete the perspective-taking task. In the example, the correct answer is 1 (Figure 1d).
The other choices are misleading random 90°, 180° and 270° rotations of the original trajectory. It also includes mirror inversions. In total, there are 4 possible positions of the red arrow. These are represented in figure 1c and denoted in table I.

[1] 3d visualization library for use with the ros javascript libraries, http://wiki.ros.org/ros3djs.

[2] https://github.com/MouadAbrini/Perspective-Taking-Assessment-Franka-Robot

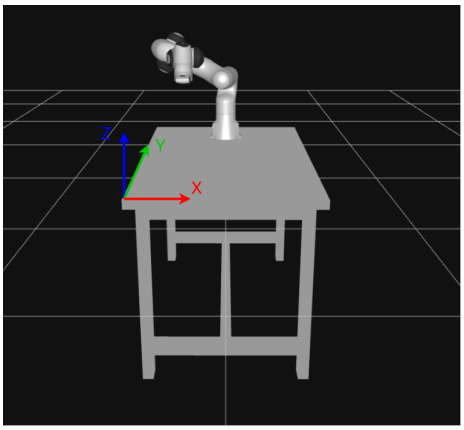


Fig. 2: *Virtual representation of the environment (egocentric view)*

Designation in figure 1c	Name	Abbreviation
0	Egocentric	P_{EGO}
1	Decentred right	P_{DR}
2	Decentred opposite	P_{DO}
3	Decentred left	P_{DL}

TABLE I: *The four possible perspectives (see figure 1c)*

2) Parameters definition: For the entire experiment, the participant had to answer N + 1 = 60 questions (excluding the example). For each question, there was only one possible correct answer. To quantify that, a binary score was given and denoted s_i ($i \in [\![0; N]\!]$ is the question number).

$$s_i = \begin{cases} 1, & \text{if the answer is correct.} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

The response time (i.e the time taken by the participant to answer at each step i) is denoted Tr_i.

3) Robot control: The robot was controlled using ROS (Robot Operating System) [27]. Two machines on which ROS noetic was installed were used for the control.

- **Computer 1:** A workstation with Ubuntu 20.04 LTS (CPU: Intel® Xeon® Silver 4214 2.2 GHz / GPU: Nvidia Geforce RTX 2080 Super)
- **Computer 2:** A desktop Ubuntu 20.04 LTS (CPU: Intel® Core™ i5-8400 2.8 GHz / GPU: Intel® UHD Graphics 630) and a patched kernel to add the real-time support required for the low-level control of the robotic arm.

The robot drawing was performed using an impedance control method with force constraint along the Z axis (see Figure 2). In fact, to ensure continuous contact with the whiteboard, we forgo the ability to move along the Z axis with translational compliance. This was necessary because the whiteboard is not a perfectly flat surface. It contains irregularities that prevent the end-effector from being always in contact with it. Hence, a force constraint has been imposed in order to compensate for these irregularities.
The Cartesian translational and rotational stiffnesses were fixed on a high enough value to ensure a low error position for an accurate drawing. The control was done using the library frankx for collaborative robots[3].

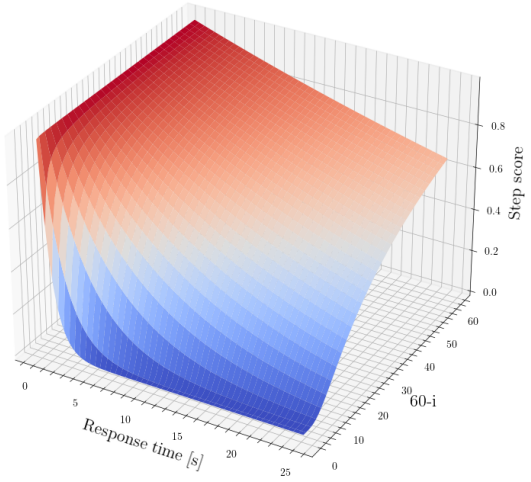


Fig. 3: *The performance metric as a function of the response time and the experiment steps, i.e., the number of completed questions*

4) Drawing speed: In order to draw the trajectory using impedance control, a set of (X,Y) pixel coordinates were collected from the user interface. These coordinates were then adapted to fit the robot's workspace. After the re-scaling of the coordinates, these were looped through and the equilibrium position of the end-effector was updated accordingly. To avoid abrupt motions of the robot, and hence to enhance the safety of the interaction, a minimum number of points was fixed under which the drawing was considered too fast. Under this threshold, the participant was asked to redraw the trajectory.

5) Metric: Previous research revealed that humans' response time decreases as they become more familiar with tasks that involve spatial perspective-taking [28]. Taking this result into consideration, we developed a metric that factors participants' response times for each question, the progress in the experiment (step number) and their answer to each question (false or correct). The metric is used to assign a score S to each participant, the score is computed following equation 2. It is calculated by averaging the values obtained from the metric at each step.

$$S = \frac{1}{N+1} \sum_{i=0}^{N} s_i e^{-\frac{Tr_i}{(N-i+1)}} \quad (2)$$

with Tr_i and s_i the response time and the binary score respectively for each step i, $N+1$ the number of questions, as defined in section III-B.2. The score S is a numerical value ranging between 0 and 1.

[3]GitHub - pantor/frankx: High-Level Motion Library for Collaborative Robots — github.com, https://github.com/pantor/frankx

The more we advance in the experiment, the more the response time will be penalized. The participant answers 60 questions (number of steps). The figure 3 illustrates the metric for N=59.

Notice the more the participant answers the questions, the steeper the curve is. So a higher response time towards the end of the trials will decrease the score significantly. But at the beginning of the experiment, the deduction for the response time is more lenient.

C. Perspective-taking and spatial orientation test (Block 2)

In this second block, the participants took the Perspective-Taking/Spatial Orientation Test (PTSOT) [29], [30]. A computerized version [4] of the test was used. The test was originally in English, but was translated into French as the participants in our study were French. The participants had 5 minutes to answer 11 questions (example excluded). The participants were provided with a reference object (flower), a direction facing another object (tree), and they were asked to point to another object (cat) as depicted in the example (Figure 4).

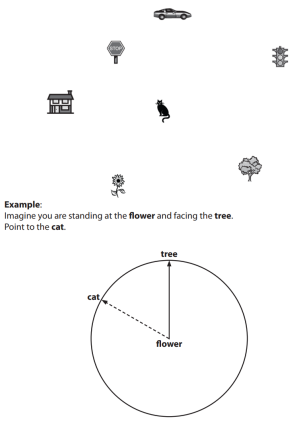


Fig. 4: *Example of the PTSOT test [29]*

Because the arrow's direction is always perpendicular to the horizontal plane, the participants had to perform spatial perspective-taking in order to find the correct answer, in this case the correct angle. An average angle error score across the 11 questions was obtained at the end of the test. It is worth noting that a lower angle error score indicates better performance in spatial perspective-taking.

IV. RESULTS AND DISCUSSION

Participants' accuracy corresponded to the proportion of correct answers on the total number of questions. 85% of the participants achieved an accuracy rate of 80 % or higher. The mean accuracy was 85 %. A cutoff accuracy of 70 % (one standard deviation away from the mean) was established to filter out the outliers. 4 participants had accuracy scores below the threshold. Two additional participants were excluded due to not following the experiment instructions. For the sequel, the results of 33 participants were analyzed (M = 28.3, SD = 6.2; 20 females, 13 males). Ages were still ranging from 19 to 42 years.

[4]GitHub - TimDomino/ptsot: Electronic version of the "Perspective Taking/Spatial Orientation Test by Hegarty, Kozhevnikov and Waller — github.com, https://github.com/TimDomino/ptsot

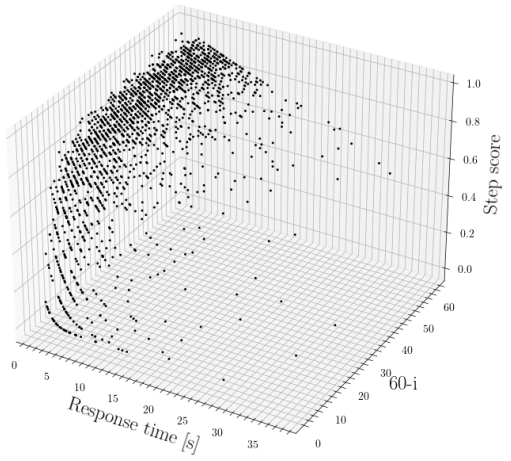


Fig. 5: *Representation of the participants' data points based on their response time at every step*

Figure 5 represents the real data based on the participants' responses, projected on the metric shown in figure 3.

A. Score computation

Figure 6 illustrates that participants' response times decrease on average as they progress in the experiment. This result is in line with the hypothesis mentioned in section III-B.5 and with previous studies on perspective-taking [28]. A quadratic least squares polynomial fit was performed to capture the evolution trend of the mean response time for each step of the experiment, considering all the participants. Since response times decrease on average as a function of the experiment steps, we can safely apply our metric.

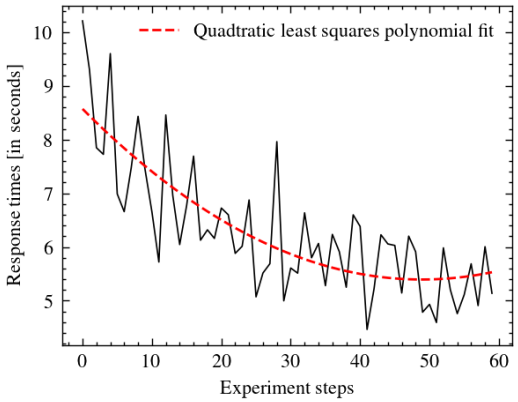


Fig. 6: *Evolution of the response time as a function of the experiment steps*

B. Correlation between the results of the first and second blocks of the experiment

The aim of the correlation analysis was to verify if there was a similar profile of performance between the first block of the experiment and the second one (the PTSOT test). This was done by creating a scatter plot and fitting a linear regression line (Figure 7).

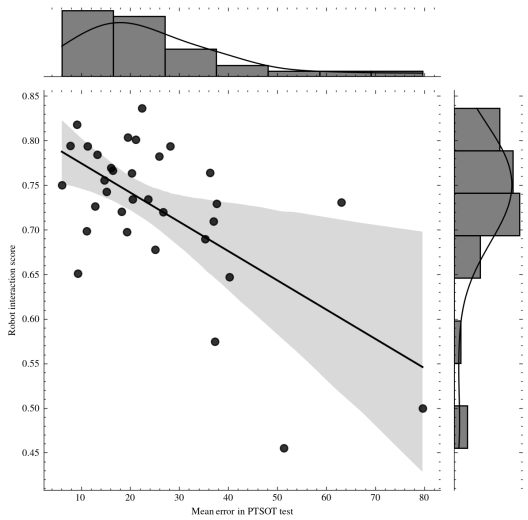


Fig. 7: *Correlation between the scores of the first block of the experiment and the angle errors of the second block*

First, Pearson's correlation [31] was computed to assess the relationship between scores of the first block and the second one. There was a significant, strong negative correlation between the two variables (r = -0.63, p = 7.55e-05). This indicates that participants with the highest scores in block 1 of the experiment have lower angle errors in block 2, as was anticipated. Our concern was that the high correlation coefficient may have been due to the few points towards the right of figure 7. In fact, this region contains very few data points, meaning that among all the participants, only few have mediocre spatial perspective-taking ability. This is supported by a study in which a computerized version of the PTSOT and a paper one were compared [32]. The angle errors distribution obtained is similar to ours. Since there was no justified reason to consider those points as outliers, we used a second correlation method. Spearman's rank correlation was computed to assess the relationship between the two scores. This method was used because it is known to be more robust to outliers and is a better use case for non normally distributed data. A moderate negative correlation between the two variables was found (r = -0.484, p = 0.003).

C. Comparison between positions

The position P_{EGO} (control condition) is the most intuitive one [28], as it does not require any spatial perspective-taking from the participants. To test whether participants' response times were faster in the control position than in the other 3 positions (P_{DR}, P_{DO}, and P_{DL}), three tests were conducted to compare the average response times between P_{EGO} (M=4.82s, SD=1.22s) and P_{DR} (M=6.86s, SD=2.9s), P_{EGO} and P_{DO} (M=6.76s, SD=3.1s) and between P_{EGO} and P_{DL} (M=6.92s, SD=2.43s). The results of the four positions is shown in figure 8. To test the normality of the data for these positions, the Shapiro-Wilk test was used. The average response times for the control condition were found to be normally distributed (W=0.97, p=0.6). This result was confirmed after an examination of the QQ plot. For the other three conditions, the data was non-normal (P_{DR}: W=0.82, p=9.54e-05; P_{DO}: W=0.86, p=5.94e-04; P_{DL}: W=0.88, p=2.43e-03). Therefore, we decided to use a non-parametric test. Since the data for the four positions were collected from the same participant, we used the Wilcoxon signed-rank test. The results revealed a significant difference between the control condition (P_{EGO}) and the other three conditions (P_{EGO}-P_{DR}: W=0.0, p=9.3e-10; P_{EGO}-P_{DO}: W=14, p=1e-07; P_{EGO}-P_{DL}: W=6, p=4.6e-09). This result shows that it is more difficult to adopt a perspective that does not correspond to our own egocentric position.

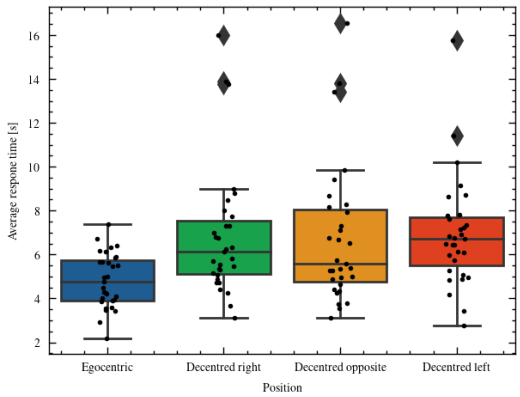
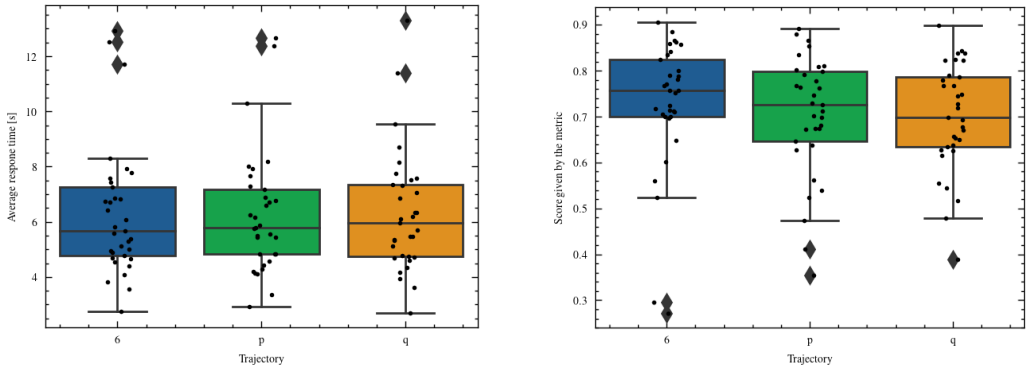


Fig. 8: *Box plot representation of the average response times for each position denoted in table I. Every datapoint (in black) in the distribution corresponds to the average response time across the 15 steps for a given participant and a position.*

Friedman test was conducted to determine whether the average response time differ between the positions P_{DR}, P_{DO} and P_{DL}. The results did not show any significant difference ($\chi^2(2) = 0.0625$, p=0.97). We therefore fail to reject the null hypothesis and conclude that there is no difference between the average response times of the three non-egocentric positions.

Even thought the statistical tests show that there was a significant difference between the average response times of P_{EGO} and P_{DO}, It is worth noting that the box plot in figure 8 indicates that adopting the decentred opposite perspective might be easier than the decentred left and right. In fact, the median of P_{DO} is slightly overlapping with the interquartile range of the control position (P_{EGO}). It is not the case for P_{DR} and P_{DL}.



(a) (b)

Fig. 9: *Comparison between the mean response time metric and our metric. Every datapoint (in black) in the distribution corresponds to the average response time or score across the 20 steps for a given participant and a trajectory. (a): Box plot representation of the mean response times for each trajectory, (b): Box plot representation of the scores computed with our metric for each trajectory*

D. Comparison between trajectories

In the same way as the comparison mentioned in IV-C, we wanted to evaluate if there was a difference between the three trajectories "p", "q" and "6". To do so, for each participant, the mean of the response times for each trajectory was computed. Then the mean of the response times of all the participants was calculated. The results are shown in figure 9a.

The results show that the response times are similar (Figure 9a) for "p" (M=6.3s, SD=2.3s), "q" (M=6.26s, SD=2.21s) and "6" (M=6.3s, SD=2.17s). A Shapiro-Wilk test revealed that the data was non-normally distributed ("p": W=0.86, p=0.00062; "q": W=0.9, p=0.004; "6": W=0.92, p=0.02). Therefore, the Friedman test was used to find out if there was a statistically significant difference between the average response times of the three trajectories. The results show non-significant difference ($\chi^2(2) = 0.41$, p=0.81). We therefore fail to reject the null hypothesis and conclude that there is no significant difference between the average response times of the three trajectories.

The three trajectories were compared in pairs using response time and our custom metric through Wilcoxon signed-rank test. The results showed no significant difference in response time data for the three possibilities (6-p: W=253, p=0.63; 6-q: W=271, p=0.87; p-q: W=263, p=0.75). The average scores for each trajectory using our metric were different, as can be seen in figure 9b. While the response time based performance measure showed little differences between the three trajectories (figure 9a), our custom metric clearly distinguished them (figure 9b). The scores are higher for trajectory "6" suggesting that observing the shape "6" from decentred perspectives may be easier compared to the other two trajectories. This is supported by figure 10 that indicates that the median response time of trajectory "6" for P_{DO} is overlapping with the IQR of P_{EGO} for the same trajectory, which is not the case for "p" and "q".

However, even using our metric, the difference between the trajectories was not significant (6-p: W=231, p=0.38; 6-q: W=207, p=0.19; p-q: W=258, p=0.7). But this may have been due to the lack of a high statistical power required to detect such a small effect size between the trajectories.

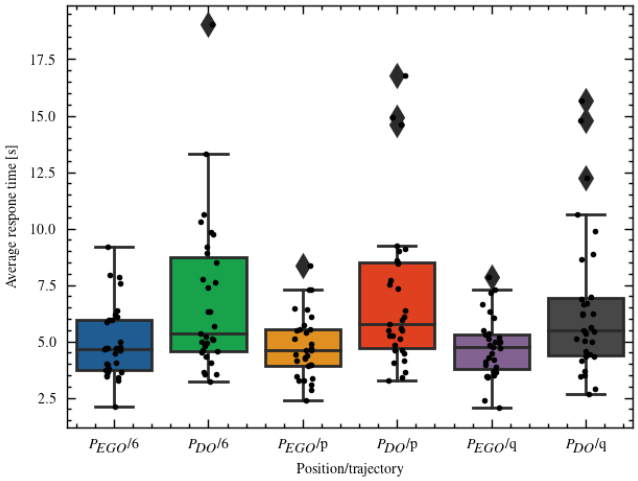


Fig. 10: *Box plot representation comparing the mean response times for P_{EGO} and P_{DO} across the three trajectories*

V. CONCLUSION

Our study highlights the importance of spatial perspective-taking in human-robot interactions and presents a novel in-person experiment to evaluate this ability. The results show that our new setup and evaluation metric are efficient in assessing perspective-taking skills in a human-robot interaction context. In addition, its validity is in line with findings from prior studies on perspective-taking. This research contributes to the development of methods to evaluate perspective-taking abilities in human-robot interactions, which is essential for the successful integration of robots into collaborative environments. By improving robots' ability to adjust their perception and movements to human agents'

abilities, we can enhance the overall effectiveness of human-robot teams. However, there are some limitations to this experiment. In particular, our method is not capable of evaluating perspective-taking ability in real-time nor adapting to dynamic changes such as fatigue. Instead, it can only generate a long-term spatial perspective-taking ability profile for the human partner. Thus, future research should investigate further assessment methods involving robots able to quickly infer humans' spatial perspective-taking abilities and able to adapt to changes in real-time.

REFERENCES

[1] P. Michelon and J. Zacks, "Two kinds of visual perspective taking," *Perception & Psychophysics*, vol. 68, no. 2, pp. 327–337, 2006.

[2] B. Tversky and B. M. Hard, "Embodied and disembodied cognition: Spatial perspective-taking," *Cognition*, vol. 110, no. 1, pp. 124–129, 2009.

[3] C. He, E. R. Chrastil, and M. Hegarty, "A new psychometric task measuring spatial perspective taking in ambulatory virtual reality," *Frontiers in Virtual Reality*, vol. 3, 2022.

[4] J. Trafton, A. Schultz, M. Bugajska, and F. Mintz, "Perspective-taking with robots: experiments and models," in *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.*, 2005, pp. 580–584.

[5] Z. S. Masangkay, K. A. McCluskey, C. W. McIntyre, J. Sims-Knight, B. E. Vaughn, and J. H. Flavell, "The Early Development of Inferences about the Visual Percepts of Others," *Child Development*, vol. 45, no. 2, pp. 357–366, 1974.

[6] J. H. Flavell, B. A. Everett, K. Croft, and E. R. Flavell, "Young children's knowledge about visual perception: Further evidence for the Level 1–Level 2 distinction," *Developmental Psychology*, vol. 17, pp. 99–103, 1981.

[7] K. Kessler and H. eRutherford, "The two forms of Visuo-Spatial Perspective Taking are differently embodied and subserve different spatial prepositions," *Frontiers in Psychology*, vol. 1, Dec. 2010.

[8] X. Job, L. Kirsch, and M. Auvray, "Spatial perspective-taking: insights from sensory impairments," *Experimental Brain Research*, Oct. 2021.

[9] A. W. Qureshi, I. A. Apperly, and D. Samson, "Executive function is necessary for perspective selection, not Level-1 visual perspective calculation: Evidence from a dual-task study of adults," *Cognition*, vol. 117, no. 2, pp. 230–236, Nov. 2010.

[10] I. Yaniv and M. Shatz, "Heuristics of Reasoning and Analogy in Children's Visual Perspective Taking," *Child Development*, vol. 61, no. 5, pp. 1491–1501, 1990.

[11] H. Moll and M. Tomasello, "Level 1 perspective-taking at 24 months of age," *British Journal of Developmental Psychology*, vol. 24, no. 3, pp. 603–613, 2006.

[12] A. Surtees, I. Apperly, and D. Samson, "The use of embodied self-rotation for visual and spatial perspective-taking," *Frontiers in Human Neuroscience*, vol. 7, 2013.

[13] S. M. Gzesh and C. F. Surber, "Visual perspective-taking skills in children," *Child Development*, vol. 56, pp. 1204–1213, 1985.

[14] X. Job, G. Arnold, L. P. Kirsch, and M. Auvray, "Vision shapes tactile spatial perspective taking," *Journal of Experimental Psychology. General*, vol. 150, no. 9, pp. 1918–1925, Sep. 2021.

[15] G. Arnold, F. R. Sarlegna, L. G. Fernandez, and M. Auvray, "Somatosensory Loss Influences the Adoption of Self-Centered Versus Decentered Perspectives," *Frontiers in Psychology*, vol. 10, 2019.

[16] X. Job, L. Kirsch, S. Inard, G. Arnold, and M. Auvray, "Spatial perspective taking is related to social intelligence and attachment style," *Personality and Individual Differences*, vol. 168, p. 109726, Jan. 2021.

[17] J. Trafton, N. Cassimatis, M. Bugajska, D. Brock, F. Mintz, and A. Schultz, "Enabling Effective Human–Robot Interaction Using Perspective-Taking in Robots," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 35, no. 4, pp. 460–470, Jul. 2005.

[18] Cassimatis and N. Louis, "Polyscheme: A Cognitive Architecture for Integrating Multiple Representation and Inference Schemes," Jan. 2002.

[19] N. L. Cassimatis, J. G. Trafton, M. D. Bugajska, and A. C. Schultz, "Integrating cognition, perception and action through mental simulation in robots," *Robotics and Autonomous Systems*, vol. 49, no. 1, pp. 13–23, Nov. 2004.

[20] H. H. Clark, *Using Language*, ser. 'Using' Linguistic Books. Cambridge University Press, 1996.

[21] A. K. Pandey and R. Alami, "Mightability maps: A perceptual level decisional framework for co-operative and competitive human-robot interaction," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 5842–5848.

[22] T. Fischer, "Perspective taking in robots: A framework and computational model," Sep. 2018.

[23] X. Zhao and B. F. Malle, "Spontaneous perspective taking toward robots: The unique impact of humanlike appearance," *Cognition*, vol. 224, p. 105076, Jul. 2022.

[24] F. I. Doğan, S. Gillet, E. J. Carter, and I. Leite, "The impact of adding perspective-taking to spatial referencing during human–robot interaction," *Robotics and Autonomous Systems*, vol. 134, p. 103654, 2020.

[25] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[26] M. Labbé and F. Michaud, "RTAB-Map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation," *Journal of Field Robotics*, vol. 36, no. 2, pp. 416–446, 2019.

[27] M. Quigley, B. Gerkey, K. Conley, J. Faust, T. Foote, J. Leibs, E. Berger, R. Wheeler, and A. Ng, "Ros: an open-source robot operating system," in *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA) Workshop on Open Source Robotics*, Kobe, Japan, May 2009.

[28] G. Arnold, C. Spence, and M. Auvray, "Taking someone else's spatial perspective: Natural stance or effortful decentring?" *Cognition*, vol. 148, pp. 27–33, Mar. 2016.

[29] M. Hegarty and D. Waller, "A dissociation between mental rotation and perspective-taking spatial abilities," *Intelligence*, vol. 32, no. 2, pp. 175–191, 2004.

[30] M. Kozhevnikov and M. Hegarty, "A dissociation between object manipulation spatial ability and spatial orientation ability," *Memory & Cognition*, vol. 29, no. 5, pp. 745–756, Jul. 2001.

[31] D. Freedman, R. Pisani, and R. Purves, "Statistics (international student edition)," *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*, 2007.

[32] A. Friedman, B. Kohler, P. Gunalp, A. P. Boone, and M. Hegarty, "A computerized spatial orientation test," *Behavior Research Methods*, vol. 52, no. 2, pp. 799–812, Apr. 2020.