Human–Robot Interaction Video Sequencing Task (HRIVST) for Robot's Behavior Legibility

Silvia Rossi, Alessia Coppola, Mariachiara Gaita, and Alessandra Rossi

***Abstract*—People's acceptance and trust in robots are a direct consequence of people's ability to infer and predict the robot's behavior. However, there is no clear consensus on how the legibility of a robot's behavior and explanations should be assessed. In this work, the construct of the Theory of Mind (i.e., the ability to attribute mental states to others) is taken into account and a computerized version of the theory of mind picture sequencing task is presented. Our tool, called the human–robot interaction (HRI) video sequencing task (HRIVST), evaluates the legibility of a robot's behavior toward humans by asking them to order short videos to form a logical sequence of the robot's actions. To validate the proposed metrics, we recruited a sample of 86 healthy subjects. Results showed that the HRIVST has good psychometric properties and is a valuable tool for assessing the legibility of robot behaviors. We also evaluated the effects of symbolic explanations, the presence of a person during the interaction, and the humanoid appearance. Results showed that the interaction condition had no effect on the legibility of the robot's behavior. In contrast, the combination of humanoid robots and explanations seems to result in a better performance of the task.**

***Index Terms*—Explainability, HRI, legibility, social robotics, theory of mind, transparency.**

I. Introduction

THE increasing use of robots in different fields of human activities requires that people accept the robots' presence and trust that they are able to complete critical tasks and look for people's well-being. In such scenarios, robots need to adapt their behaviors to overcome the possible negative attitudes of people toward them (e.g., fear and anxiety), which may negatively affect people's trust in robots [1]. To address these issues, the relatively new research area called explainable robotics began to study explainability in human–robot interaction (HRI) [2]. This field investigates how to make a robot's behavior legible ([3], [4]) by providing explanations of its actions and recommendations. Existing works on explanation evaluation in AI usually break down the problem into the evaluation of narrower aspects of an explanation, such as its intelligibility, accuracy, precision, completeness, and consistency (see [5], [6] for a survey on the state of the art). These criteria are generally context-dependent and hard to assess in a general way. They typically do not consider the role of physical actions and the effects of human–robot collaborations on the robot's behavior legibility.

Our purpose, instead, is to design context-independent evaluation metrics starting from the assumption that one of the main drives to generate explanations and develop legible behavior in HRI is to easily communicate the robot's goal or "intent" independently of the specific application scenario [7]. We also believe that the correct attribution of a robot's goal-directed behaviors depends on its ability to perform actions that are readable, and to provide the correct explanations when it is needed. The ability to attribute mental states, such as beliefs, goals, and intentions different from one's own [8], and the ability to understand and predict the behaviors of others [9], are indispensable processes to effective social interactions, that are known as Theory of Mind (ToM).

Is mind attribution necessary to correctly understand robots' behavior? Do people need to think that a robot can have a rational intention to be able to predict what the robot is going to do? According to some authors ([10], [11]), the answer to these questions may be "yes." The influence of people's mental representations of robots in predicting their behaviors seems to be shown by [11], who observed people forming a mental model of the robot's knowledge based on their own knowledge and on the information about the robot's origin and language. Huang et al. [10] highlighted the inadequacy of goal understanding to effectively predict a robot's future behavior, and emphasize the importance of having a good mental model of the robot's decision mechanism. According to the authors, people need to have an implicit representation of what drives the robot's behavior (i.e., a qualitative understanding of the tradeoffs the robot makes to achieve a certain goal). A similar distinction can be found also in Dragan et al. [12] that differentiates legibility, which is defined as the ability to correctly infer the goal and predictability, which is defined as the ability to predict the specific action that will achieve that goal. Hellström and Bensch [13] instead introduced the term "understandability" defined as having sufficient knowledge of the robot's state of mind in order to successfully interact with it. A similar definition of legibility is provided by Takayama et al. [14] according to

Manuscript received 6 July 2023; revised 11 September 2023; accepted 20 October 2023. Date of publication 14 November 2023; date of current version 14 December 2023. This work was supported in part by Italian MUR and by the European Union's CHIST-ERA project COHERENT under Grant PCI2020-120718-2 (CUP E63C21000100006), and in part by Italian PON R&I 2014-2020—REACT-EU under Grant CUP E65F21002920003. This article was recommended by Associate Editor J. A. Adams. *(Corresponding author: Silvia Rossi.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethical Committee of Psychological Research of the University of Naples Federico II under Application No. 13/2022, and performed in line with the ethical standards of the 1964 Declaration of Helsinki.

The authors are with ICAROS Center, The University of Naples Federico II, 80131 Napoli, Italy (e-mail: silvia.rossi,@unina.it; alessiacoppola001@gmail.com; mariachiara733@gmail.com; alessandra.rossi@unina.it).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/THMS.2023.3327132.

Digital Object Identifier 10.1109/THMS.2023.3327132

which the robot behavior is legible if people can figure out what the robot is doing, and reasonably predict what the robot will do next, and ultimately effectively interact with the robot. These definitions show a partial overlapping between legibility, as defined in human–robot interaction, and ToM, as defined in human–human interaction. In this direction, we believe that a comprehensive definition of legibility is well provided by Lichtenthaler and Kirsch [15], and it is defined through two factors: 1) understandability of intentions (an umbrella term including predictable action, predictable motion trajectory, and predictable goal); and 2) the matching between the behavior of the robot and the expectations of the human observer or interactor.

In this regard, the aims of the present study are: 1) to create a valid measurement tool that assesses the ability of humans to attribute a goal-oriented behavior to robots and to correctly read their behaviors in different contexts (e.g., both in the case of humanoid and nonhumanoid robots or if they are interacting with a user or performing actions on their own); and 2) to start assessing the role of symbolic explanations in improving human prediction of robots' behavior. We created a variation of the theory of mind picture sequencing task, which could be applied to social cognition in human–robot interaction. The resulting tool, called the human–robot interaction video sequencing task (HRIVST) involves watching a randomly ordered video sequence of actions while trying to put them in the correct logical order. The test is accompanied by two open-ended questions assessing the understanding of the intentions of the robots. We recruited a sample of 86 healthy subjects. Results showed that the HRIVST has good psychometric properties and could be a valuable tool for assessing the legibility of the robot's behavior in terms of the human ability to solve the task and attribute a goal-oriented behavior to the robot. Results also showed that the interaction condition has no effect on the legibility of the robot's behavior, while better performances were obtained in videos presenting humanoid robots and no explanations.

II. Background and Related Works

In this work, we start from the hypothesis that the need for legibility and explainability of a robot's behavior is a direct consequence of the need to understand what the robot is doing or is going to do. Since the ToM has been assumed to be a unitary process supporting different functions depending on which kind of judgments are made (on intentions, emotions, beliefs, etc.), there is a wide variety of measurements to assess this ability ([16], [17]). The measurements are different for test formats (verbal, visual, audiovisual, or scale) and aims.

A. ToM Evaluation in Human–Robot Interaction and Association With Mind Attribution

In human–human interaction, ToM is an automatic and natural process that allows people to understand that others have mental states, such as preferences, beliefs, desires, and intentions that are different from their own [8]. However, there is no clear consensus on whether and when mind attribution can be elicited when a human is dealing with a robot.

The propensity to recognize nonhuman agents as capable of having a "mind" can vary depending both on the agent's external features (e.g., appearance and behavior) and the observer's internal dispositions, such as beliefs, expectations, motivations, individual differences, and experience ([18], [19]). As for the agent features, a relevant role in mind attribution is played by appearance and behavior. For example, it is possible to induce the adoption of a mental state toward a humanoid robot, due to its human-like appearance [20]. A human-like robot is also more likely to elicit anthropomorphic interaction and attributions and to be considered more life-like and sociable than an artificial agent created from a photo of the same robot's head and neck [21]. Specifically, the study of face perception through the evaluation of images of animate (human) to inanimate (mannequin) faces, varying along a spectrum, shows that life and mind attribution appear only after a categorical threshold, close to the human endpoint [22]. Similarly, physical manipulation of agent appearance, ranging from mechanical to humanoid and from humanoid to human, can influence mind attribution in human observers. After a certain threshold, increasing human-like appearance can modulate the attribution of specific mental states to an agent (e.g., feeling pain, being able to move, etc.) [18].

Neuroimaging studies detected the same neural circuitry underlying the perception of human and anthropomorphized robot behavior [23]. The authors showed that the degree of anthropomorphism of an agent modulates cortical activity in previously detected ToM networks. They showed that participants interacting with a robotic partner had higher activity modulation than when interacting with a computer [24].

Regarding the agent's behavior, mentalization processes can be triggered if they show movements reminiscent of human social interaction despite the anthropomorphism of the robot [25]. Mental states can be attributed to both human and nonhuman agents (e.g., toy robots) when they move at speeds that approximate normal human motion, but they appear mindless when they move either faster or slower than normal human motion [26]. Moreover, mentalizing processes are consistent across different human-like robots and humans when displaying social cues (e.g., eye movements, voice, appearance) that are similar and interpretable. In particular, performances to divergent experience recognition tasks (e.g., false belief test) are similar when the characters displayed are humans or robots, independently of social cueing provided (e.g., round body robot with squeak noises versus bipedal humanoid robot with human-like voice) suggesting that mind attribution can be elicited when robot behaviors follow human scripts [27]. This finding seems to be in contrast with a study that showed that fewer participants passed a false belief test when the protagonist was a humanoid robot rather than a human [28].

We can note that external factors alone are not able to account for mind attribution to robotic agents. The same behavior can be interpreted in different ways according to the observer's beliefs, and it can modulate the observer's behavior. For example, people are more willing to follow the gaze direction of a robot under the assumption that its eye movements are controlled by a person, rather than preprogrammed, proving that manipulating high-level cognitive processes, such as the likelihood of adopting the intentional stance, can modulate bottom-up attentional

mechanisms [29]. The attribution of human-like intentions to nonhuman agents can also be influenced by psychological and social factors, such as the need to effectively interact with the environment, to satisfy the typical human desire to establish social connections or to reduce the anxiety deriving from the uncertainty in anticipating others' behaviors. Thus, a more likely tendency to anthropomorphize nonhuman agents should be expected when the motivation to feel efficacious increases or when people feel a lack of social connection in the absence of other humans [30]. This also suggests an important role of personality traits (e.g., the tendency to feel socially disconnected or the need for control) in mind attribution [31]. The anthropomorphic aspect of a humanoid robot can modulate the observer's expectations [32]. For instance, although human appearance has a positive impact on first-encounter attitudes, its effect seems to have negative consequences on HRI if the expectation that a robot behaves as a human as induced by a robot's appearance is not met by its actual behavior [33].

Since there is no clarity on which conditions may elicit mind attribution to robots, we decided to investigate this issue by testing the proposed tool on a variety of robots' goal-oriented behaviors and comparing our measure with a false belief's task to gain insight into its psychometric properties.

B. ToM Measurements in Human–Human Interaction

A preliminary analysis of measures evaluating cognitive ToM has been conducted to create a theoretical basis for constructing a new metric that could appropriately assess human understanding of robot behavior. Among the different measures, a great value is recognized in those that, besides other features, reduce the influence of verbal deficits and possess greater ecological validity [34]. These use materials like short stories or picture stories, in which two or more characters interact and the subject's task is to recognize that the character's belief does not match shared world knowledge or reality (false belief's task). In addition, some assessment materials contain control tasks requiring mechanical cause-and-effect reasoning [35] or reality questions [16] to control interferences with memory, attention, and verbalization.

One of the widely used measures to assess the ability to attribute cognitive states to others, in both verbal and nonverbal modalities, is the ToM Picture Sequencing Task (TPST) [16]. It consists of the presentation of six cartoon picture stories of four cards each, depicting two scenarios where two characters cooperate, two scenarios where one character deceives a second character, and two scenarios where two characters cooperate in deceiving a third. The cards are presented in a mixed order and participants are asked to arrange them in a logical sequence of events. The logic reasoning task is followed by 23 questions investigating comprehension of first, second, and third-order beliefs, understanding of the cheating, and comprehension of reciprocity.

Although false belief tasks like TPST are considered the gold standard test in evaluating cognitive ToM [36], the use of static images limits ecological validity. Attempts to overcome this limit in ToM assessment have led to the development of video-based tasks detecting the understanding of social interactions and behaviors based on the observation of dynamic over static stimuli, depicting real people in real-life contexts. Video-based tasks have been created from existing excerpts of films [37], with the quality of an actual movie [38], or vignettes with professional actors and a camera team [39].

Clips from television commercials and series showing a character experiencing a socially awkward or unpleasant moment were used in the awkward moments test. The test comprised of questions referring to a character's feelings besides control questions about a visual or verbal feature within the film or the dialog, serving as general indices of attentional processing and memory for information [37]. Relevant features of the TV commercials' selection were their high-quality technical production, short duration (from 45 to 120 s), complete storyline, no need for prior knowledge of the characters, variety of situation settings and characters' age, roles, and relationships.

Despite their capacity to approximate everyday social interactions more adequately than static pictures or story formats, video-based tasks strongly involve executive functions and central coherence. To minimize demands on these cognitive functions and gain more control over the generation of mental states to be inferred, Dziobek et al. [38] created the Movie for the Assessment of Social Cognition (MASC) by shooting a 15-min film made from scratch, avoiding distracting or prompting stimuli such as music, direct camerawork, a complete storyline, and fast-changing scenes that are typical of TV clips. Respondents' task consists of watching the movie depicting everyday social interactions and answering questions about the actor's mental states (thoughts, emotions, intentions).

III. Human–Robot Interaction Video Sequencing Task

On the basis of the background and related works so far explored, we designed a tool to assess the legibility of a robot's behavior by evaluating the human's ability to properly interpret the intentions of the robotic agent, and therefore of its final goal, by looking at a sequence of goal-oriented actions. To evaluate this, we proposed a new version of the well-known TPST test that could be applied to the evaluation of social cognition in HRI. Our solution differentiates from the TPST in the use of video clips instead of pictures that the subject has to order correctly to create a logical sequence of events. We believe that watching video clips of a robot performing actions may contribute to greater ecological validity.

To apply the metrics to the evaluation of a robot's behavior, such behavior has to be recorded in a video. The video showing a goal-directed behavior has to be divided into several video clips, one for each single, purpose-directed, robot action so that humans could have temporal projections to create a logical sequence. Indeed, cognitive psychology showed that people perform their actions based on intended consequences [40], and they tend to perform mental simulations through mirror neurons to understand the final outcome [41].

Then, the test consists of asking the participants to order the video clips to form a sequence of logical sense in the shortest time possible. As in TPST, we measure people's sequencing ability using the scoring method of Langdon et al. [35]. Therefore, we assign two points each for both the first and the last

video clips, while we assign two points divided by the number of remaining clips if people correctly order the remaining clips. For example, in the case of a video composed of four clips, four points are assigned for the first and last video clip, and one point is assigned for correctly positioning each of the central clips within the sequence. Thus, for each correct sequence, the subject can obtain a maximum score of six points.

To measure whether people are able to recognize the robot's actions as intelligible and understandable, participants underwent a semistructured questionnaire that had a twofold task. We used it to investigate what was objectively the action that the robot was performing ("What is the intention of the robot?") and, similarly to the TPST, we wanted to measure the robot's predictability ("What do you expect the robot to do next?"). This first phase of the questionnaire allowed us to investigate the correct understanding of the robot's intentions and the attribution of mental states. We assigned a one-point score for each correct answer. A high and low score indicate respectively a good and bad understanding of the intentions of the robots. We then assessed the subject' perceived confidence in correct intentionality attribution ("How difficult was it for you to attribute intention to the robot?") using a Likert scale [from 1 to 5]. Thus, the total score of our measuring instrument is given by the sum of the score obtained from the logical sequence task (i.e., 0 to 6) and the score of correct answers given to the questionnaire (i.e., 0 to 2), so that the total ranges from 0 to 8, for each video.

IV. HRIVST Validation

The task validation consisted of three phases: 1) video selection; 2) video clip creation; and 3) testing. Moreover, we decided to address the effect of some factors on human understanding of the robot's behavior, with specific reference to: the humanoid appearance of the robot; the presence of symbolic explanations. We also investigated whether the presence of a human agent involved in the robot's actions may influence the capacity to predict its behavior.

A. Video Selection

In the first phase, we selected 12 videos from YouTube[1] showing humanoid (**humanoid** condition) or nonhumanoid robots (**nonhumanoid** condition) performing sequences of actions on their own (**noninteraction** condition) or in interaction with a human (**interaction** condition). The selection has been conducted following the same criteria used in the literature to minimize distracting effects deriving from fast-changing scenes or music [38].

To gain greater ecological validity, we chose videos showing different tasks but with a similar duration, with a range between 12 and 48 s. For the majority of the videos, we decided to include those with robotics tasks, which are common to industrial and service robots, such as delivery tasks that also involve manipulation [42].

[1]The list of the used videos and the action samplings is available on request.

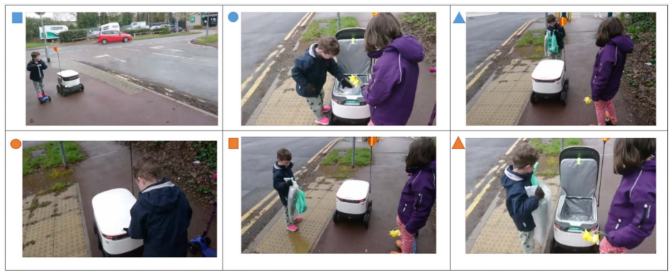

Fig. 1. Video 5 from Table I is an example of the interaction of a nonhumanoid robot. It was used for HRIVST validation.

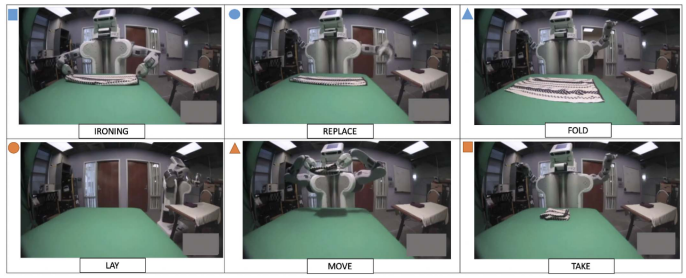

Fig. 2. Example of a noninteracting video of the HRIVST test. Video 11 shows a humanoid robot doing a folding task.

B. Video Clips Creation

The videos were divided into several video clips, with the number of clips varying from three to six. We considered a video composed of three or four clips as having a "short length"; videos of five clips as having a "medium length"; and videos of six clips as having a "long length." Furthermore, since knowing the behavior of a robot improves the understanding of the actions [43], some video clips were modified to feature simple explanations in terms of keywords. The explanations were provided during the robot's action. Explanations are to be considered as communicative components that create a manageable and shared meaning, which influences the mind or behavior of those who observe the actions performed by the robot [42]. Since people always seek explanations to clarify and better understand intentions, some of the selected video clips were left without explanations to assess whether, in the absence of verbal clues, the robot's intent was equally clear, and to verify whether this might have consequences on the robot's behavior prediction accuracy. The 12 videos used to validate the HRIVST measure are described in Table I.

C. Testing

With HRIVST, we want to create a subjective measure to assess a robot's behavior legibility through the evaluation of people's ability to attribute goal-oriented behavior to the robot. To validate such a measure, we administered the selected 12 videos to the participants. Figs. 1 and 2 show examples captured from two videos used in the validation study. Both videos were divided into six video clips, one for each action done by the robot or the people involved in the interaction (i.e., two children in Fig. 1). Participants were asked to watch each video clip and order them according to the temporal sequence of the actions.

TABLE I
FEATURES' DESCRIPTION OF THE VIDEOS USED TO VALIDATE THE HRIVST QUESTIONNAIRE

Video#	Video Description	Interaction	Humanoid	Explanations	Length
Video1	A robot opens a bottle and pours the liquid inside into a paper cup, then gives it to a man sitting in an armchair and reading a book	yes	yes	yes	medium
Video2	A robot takes a tool and brings it to a human who is assembling an object in an industrial-like environment	yes	no	yes	long
Video3	A robot holding a tray with bottles navigates and stops next to a girl sitting at a table. The girl picks the bottles from the tray and puts them on the table. The robot turns away and navigates back	yes	yes	yes	medium
Video4	A robot picks a little ball off the floor and gives it to a girl who's watching him and smiling	yes	yes	no	short
Video5	A robot navigates along a street. A kid follows it, opens the robot when it stops and picks some full shopping bags from the inside with the help of a little girl, then the kids wait for the robot to leave	yes	no	no	long
Video6	A robot picks a fork from the wheelchair's side of a boy sitting at a table, then forks a piece of fruit from a plate on the table and approaches the boy's mouth, who eats it	yes	no	no	medium
Video7	A robot picks one of three origami off a room floor and puts it in a basket	no	yes	no	short
Video8	A robotic arm opens a drawer, picks up a toy machine from a wooden platform and puts it into the drawer, then it approaches a second drawer	no	no	no	medium
Video9	A robot picks an ice cream cone from a holder and fills it with two scoops of ice cream of different flavours; then it puts the ice cream cone on support	no	no	no	long
Video10	A two arms robot grasps a bottle's cork and twists it	no	no	yes	short
Video11	A robot folds a cloth many times on a table then picks it up and puts it on another table	no	yes	yes	long
Video12	A humanoid robot picks the pieces constituting a vacuum cleaner up, assembles them together and shows how it is used	no	yes	yes	short

They were allowed to watch the clips as many times as they wanted. The geometric shapes associated with each video clip need to be named in the correct order to create the sequence. If participants selected the wrong sequence, the examiner showed them the correct sequence of actions. For example, the correct sequence of actions, as shown in Fig. 1 is given by the following sequence: (blue square) the robot navigating next to a child; (orange circle) a child opens the robot's trunk; (blue circle) the child picks some shopping bags up with the help of the other child; (orange triangle) the first child closing the robot's trunk; (orange square) the kids waiting for the robot to leave; (blue triangle) the robot leaves. Similarly, Fig. 2 shows a video in which a humanoid robot is folding scarfs with no human interaction.

After each video sequence, participants were asked to answer a short questionnaire and identify the intention of the robot's actions, their expectation of the robot's actions, and their confidence in attributing intention to the robot (i.e., it was difficult or easy). The first two questions were tailored to reflect the video sequence. For example, related to the video in Fig. 1, they were asked to answer the following set of questions: 1) **intention**: What is the intention of the robot? [Correct answer: bring a shopping bag/something]; 2) **expectation**: What do you expect the robot will do next? [Correct answer: go away/get away]; and 3) **confidence**: How difficult was it for you to attribute intention to the robot? [Scale from 1 to 5, where 1 is "not at all" and 5 is "very"].

D. Participants

One hundred healthy subjects (HCs) were recruited and screened for eligibility. Prior to their participation in the study, people underwent a brief neuropsychological assessment using the Mini-Mental State Examination (MMSE) [44] and Raven's Colored Progressive Matrices [45] tests. The MMSE test consists of a 30-point questionnaire that takes from 5 to 10 min, and it examines people's functionalities, including orientation, attention, memory, language, and visual-spatial skills. The Raven's Colored Progressive Matrices [45] is a nonverbal test typically used to measure general intelligence and abstract reasoning.

TABLE II
DESCRIPTIVE STATISTICS OF DEMOGRAPHIC AND COGNITIVE VARIABLES, WHERE MMSE IS THE MINI-MENTAL STATE EXAMINATION, AND RCPM IS THE RAVEN'S COLORED PROGRESSIVE MATRICES

	Demographic Variables		*Cognitive Variables*	
	Age	Education	MMSE	RCPM
Mean (SD)	28.64 (8.55)	15.79 (2.05)	29.65 (0.61)	31.93 (2.26)
Min - Max	18-60	13-19	28-30	28-36

After this prescreening, we included only 86 participants based on the following criteria: no evidence for dementia or cognitive decline as defined by age, and education-adjusted score on the Mini-Mental State Examination lower than 23.8 according to Italian norms [46]; no deficits in abstract reasoning as defined by age and education adjusted score on the Raven's Colored Progressive Matrices [47]; no psychiatric or neurological disorders as assessed through the question "Have you ever suffered from psychiatric or neurological disorders?"

Participants' demographics (e.g., age and education in years) and cognitive variables (MMSE and RCPM total scores) are reported in Table II.

The study was performed in accordance with the ethical standards of the 1964 Declaration of Helsinki. It was approved by the Ethical Committee of Psychological Research of the University of Naples Federico II (n.13/2022). People also gave written informed consent prior to their participation.

E. Assessing Participants' ToM Ability

To test the ability of our tool to assess the attribution of intentions, it was necessary to compare it with a widely used questionnaire with good reliability. For this reason, a computerized version of the TPST was administered to participants to assess their ability to infer others' mental states. We adopted the scoring method of Langdon [35] for measuring people's ability to sequence the images, which assigns two points each for the first and the last picture, and one point for each middle picture, if they are correctly sequenced. The maximum score is

TABLE III
DESCRIPTIVE STATISTICS OF THE HRIVST

HRIVST (n= 86 healthy subject)					
	Mean (SD)	Min-Max		Mean (SD)	Min-Max
Video1	6.63 (1.81)	0-8	Video2	4.72 (1.57)	0-6
Video3	6.72 (1.47)	2-8	Video4	7.79 (0.69)	5-8
Video5	6.08 (2.01)	1-8	Video6	7.11 (1.22)	3-8
Video7	7.92 (0.27)	7-8	Video8	5.76 (1.97)	1-8
Video9	6.57 (1.66)	2-8	Video10	7.50 (1.33)	1-8
Video11	6.34 (1.38)	3-8	Video12	6.34 (1.38)	1-8

six points per story, with an overall score of 36 points for the six stories. In the case of incorrect sequencing, the rater corrects the sequence before asking questions. For each correct answer, 1 point is given, with an overall score of 23 points. The total score of the TPST ranges from 0 to 59.

F. Statistical Analysis

An a priori sample size calculation using G*Power 3.1.9.2 considering ANOVA tests within factors, as analysis, was conducted. We used an Effect size $f = 0.25$ for a large effect size, an error probability equal to 0.05, Power ($1 - \beta$ error probability) equal to 0.95, and $df = 2$, which reflected the number of fixed choices in our measure, and a number of measurements equal to 4, representing the number of the independent conditions in our study. The sample size calculation for a sufficiently powered statistical analysis resulted in 86 HC participants, which is exactly the number of participants remaining after the prescreening assessment. We analyzed psychometric properties by exploring internal consistency and construct validity and considered the level of internal consistency Cronbach's α as an acceptable level when its value is equal to or greater than 0.70 [48].

The construct validity was evaluated through convergent and discriminant validity. The convergent validity was assessed by correlation analysis between the HRIVST total score of each video and the TPST total scores of each short story. The discriminant validity was assessed by correlation analysis between the HRIVST, RCPM, and MMSE scores. To evaluate and compare the performance of the participants in the HRIVST test, we used an ANOVA test with dichotomization of the following conditions:

1) *interaction* analyzing the differences in the scores for videos with human interaction and noninteraction;
2) *humanoid* analyzing the differences in the scores for videos with humanoid and nonhumanoid robots;
3) *explanation* analyzing the differences in the scores for videos with explanation and nonexplanation;
4) *length* analyzing the differences in the scores for videos with different length (short, medium, long).

The HRIVST total scores for each video were used as dependent variables.

Analyses were followed up by Bonferroni-adjusted post hoc tests for further investigating whether, independently of the examined conditions (interaction, humanoid appearance, explanation, and length) some videos could be more difficult to assess for the participants.

V. RESULTS

The mean values and SD obtained for each video of the questionnaire are reported in Table III.

A. Quality of Data and Internal Consistency

A Kolmogorov–Smirnov test was used to verify the normal distribution of the data. Since the results and histogram indicated that the data were normally distributed, parametric analyses were performed.

The HRIVST had high internal consistency (Cronbach's $\alpha = 0.970$) showing a good degree of agreement between the participants' answers given to each video of the questionnaire.

B. Construct Validity

The HRIVST total score correlated moderately and significantly with the TPST total score ($r = 0.329$, $p = 0.004$), but not with the MMSE ($r = 0.036$, $p = 0.744$) and RCPM ($r = -0.014$, $p = 0.902$). This result showed that the HRIVST questionnaire can be considered a reliable tool to assess the correct understanding of the robot's intentions by humans. Bonferroni's correction showed that the second short figure story of the ToM task was moderately and significantly correlated with Video 2 ($r = 0.377$, $p < 0.001$); and Video 3 ($r = 0.339$, $p = 0.001$); Video 10 ($r = 0.377$, $p \leq 0.002$); Video 12 ($r = 0.337$, $p = 0.002$) of the HRIVST questionnaire. The third short-figure story of the ToM task was moderately and significantly correlated with Video 2 ($r = 0.319$, $p = 0.003$) and Video 10 ($r = 0.372$, $p = 0.001$). While the fifth short figure story of the ToM task was moderately and significantly correlated with Video 1 ($r = 0.350$, $p = 0.001$); Video 2 ($r = 0.334$, $p = 0.002$) and Video 10 ($r = 0.418$, $p \leq 0.001$). Correlations between each HRIVST video and the TPST stories are reported in Table IV.

C. Effect of Anthropomorphism, Interaction, and Explanations

An ANOVA test comparing the total scores obtained by the participants to each video on the basis of the different conditions showed: 1) no statistically significant effect for the **interaction** condition ($F = 0.451$; $p = 0.652$), meaning that the presence or absence of a human interacting with the robot in the video does not influence the participant's answers; 2) a statistically significant effect for **humanoid** condition ($F = 3.911$; $p = 0.048$), and in particular, participants showed higher total scores to videos with humanoid robots than with nonhumanoid robots; and 3) a statistically significant effect for the **explanation** condition ($F = 5.433$; $p = 0.020$), with participants showing higher total scores to the videos that did not present the explanations.

To assess the effect of explanation in relation to the robot's type, we conducted an ANOVA with the total scores obtained from the videos only featuring nonhumanoid robots in the **explanation** condition (with or without explanations). The results showed that the participants have a higher score for the videos that presented explanations ($F = 0.028$; $p = 0.866$), even though this effect was not statistically significant. Finally, an analysis of the effect of the **length** condition (short, medium, long sequence) was statistically significant ($F = 22.856$; $p \leq 0.001$). In particular, these results showed that people had

TABLE IV
CONVERGENT AND DISCRIMINANT VALIDITY OF THE HRIVST QUESTIONNAIRE WITH MMSE STANDS FOR MINI-MENTAL STATE EXAMINATION, RCPM FOR RAVEN'S COLORED PROGRESSIVE MATRICES, AND SF FOR SHORT FIGURE STORY OF PICTURE SEQUENCE TASK

	V1	**V2**	**V3**	**V4**	**V5**	**V6**	**V7**	**V8**	**V9**	**V10**	**V11**	**V12**
MMSE	-0.085	-0.002	0.135	0.133	0.012	0.126	0.068	-0.142	0.047	0.014	0.188	-0.121
RCPM	0.033	-0.099	0.166	-0.055	-0.209	0.288	-0.11	-0.111	0.249	-0.128	0.126	-0.203
SF1	0.136	0.218	-0.131	0.129	0.073	-0.141	-0.096	0.286	-0.131	0.081	0.01	0.216
SF2	0.227	**0.377**	**0.339**	0.038	0.14	-0.236	0.07	0.3	-0.146	**0.377**	0.067	**0.337**
SF3	-0.091	**0.319**	0.063	0.066	0.026	-0.188	0.198	0.056	0.002	**0.372**	0.064	0.109
SF4	-0.07	0.201	-0.074	-0.026	-0.099	-0.062	-0.023	-0.071	-0.073	-0.032	-0.009	-0.125
SF5	**0.350**	**0.334**	-0.279	-0.076	-0.036	-0.239	-0.021	0.283	-0.258	**0.418**	0.14	0.295
SF6	0.207	0.131	-0.098	0.008	-0.141	-0.115	0.249	0.04	0.027	0.226	0.093	0.14

The value with p 0.002 after bonferroni correction are reported in bold.

TABLE V
COMPARISON OF THE PARTICIPANTS' PERFORMANCE IN EACH VIDEO

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12
V1	-	**1.910***	-0.087	-0.259	0.551	-0.477	**-1.363°**	**0.877***	0.067	**-1.363***	0.288	**1.590***
V2	**-1.910°**	-	**-1.996°**	-2.373	**-1.359°**	**-2.387°**	**-3.273°**	**-1.033°**	**-1.843°**	**-3.273°**	**-1.622°**	-0 .320
V3	0.087	**1.996***	-	-0.172	0.637	-0.391	**-1.277°**	0.963	0.153	**-1.277°**	0.374	**1.677***
V4	0.259	2.169	0.172	-	0.809	-0.219	-1.105	1.135	0.326	-1.105	0.547	1.849
V5	-0.551	**1.359***	-0.637	-0.809	-	**-1.028°**	**-1.914°**	0.326	-0.484	**-1.914°**	-0.263	**1.039***
V6	0.477	**2.387***	0.391	0.219	**1.028***	-	**-0.886°**	**1.354***	0.544	**-0.887°**	**0.765***	**2.067***
V7	**1.363***	**3.273***	**1.277***	1.105	**1.914***	**0.886***	-	**2.240***	**1.430***	0.001	**1.651***	**2.953***
V8	**-0.877°**	**1.033***	**-0.963°**	**-1.135°**	-0.326	**-1.354°**	**-2.240°**	-	**-0.810°**	**-2.240°**	-0.589	0.713
V9	-0.067	**1.843***	-0.153	-0.326	0.484	-0.544	**-1.430°**	**0.810***	-	**-1.430°**	0.221	**1.523***
V10	**1.363***	**3.273***	**1.277***	1.105	**1.914***	**0.887***	0.001	**2.240***	**1.430***	-	**1.651***	**2.953***
V11	-0.288	**1.622***	-0.374	-0.547	0.263	**-0.765°**	**-1.651°**	0.589	-0.221	**-1.651°**	-	**1.302***
V12	**-1.590°**	0.320	**-1.677°**	-1.849	**-1.039°**	**-2.067°**	**-2.953°**	0.713	**-1.523°**	**-2.953°**	**-1.302°**	-

The values marked with "* " indicate the statistically significant worse score. The Bonferroni correction significant threshold set at p 0.05 is reported in bold.

a higher score for the videos that presented a short sequence compared to the long and medium-long sequence, and the participants scored worse in the medium sequence than in the long sequence.

D. Comparison Between the Videos HRIVST Scores

A Bonferroni-adjusted post-hoc analysis showed some significant differences in the scores participants obtained for the items of the questionnaire for each video. The results with statistically significant differences for each individual video are presented in Table V.

VI. DISCUSSION

The results of this study showed that the HRIVST test has good psychometric properties, in terms of reliability and construct validity (both convergent and discriminant), providing a valuable metric for future investigations. The correlation between the HRIVST and the TPST scores indicated a fair convergent validity, and, therefore, suggests that our test is able to assess people's ability to attribute mental states. The HRIVST also has good internal consistency, with a Cronbach's α value of 0.97. Observing that participants had similar HRIVST scores for each video, this result leads us to believe that the difference in the video average results is linked to the legibility of the robot's behaviors across different tasks. The low and not significant correlation between the HRIVST score and cognitive tasks scores (MMSE and RPCM) indicated good discriminant validity, meaning that the considered tasks assess different constructs. Building on these preliminary data based on healthy participants, we highlight the need for future studies to investigate the effect of different age groups of adults on the ability to attribute mental states to robots.

In particular, the results of our study highlighted that:

a) Individual participants' scores significantly differ depending on each video: The comparison of the mean scores for each video showed some remarkable patterns: The worst scores were obtained in Video 2 and Video 8, while the best score was obtained in Video 7. In Video 2 participants showed significantly worse scoring than all videos except for Video 4 and Video 12. Indeed, Video 2 presented all the conditions resulting in lower scores, such as the presence of explanations, the longest sequence, and the presence of a nonhumanoid robot. In Video 2, deficient performances may also be due to the used characteristics for creating the video, such as viewpoint, lighting, and frame resolution. These seem to have worsened the resulting scores in the nonhumanoid, the long sequence, and the explanation conditions. In Video 8, participants may have scored lower in the nonhumanoid condition than in the humanoid condition for similar reasons. Indeed, action recognition through video may be difficult when the video's recording quality does not provide a full overview of the scenario [49]. In a complex sequence of actions where the single movements cannot always be clearly distinguished, the detection of objects may help to a better understanding of the ongoing event, providing useful information [50]. This may not be the case in Video 8, in which a robot opens a drawer, grasps a toy car from one of the shelves, puts the car in the drawer, closes it and opens another drawer. In this video, multiple cars are visible, and this may have led the participants to confuse the opening and closing actions of the drawer. As a result, the sequence and the robot's goal could have been more difficult to discern, resulting in the worst scores. On the contrary, Video 7

resulted in having the highest score with no explanation, a short length of time, and a humanoid robot. In summary, we found that the significant differences in participant scores for each video are consistent with the existing literature about the methods to achieve legible robot behaviors [15]. In particular, to generate legible robot behaviors, the following variables were proposed: human-like behavior ([51], [52]); stereotypical motions of everyday activities (e.g., reaching for an object) ([51], [53]); the addition of complementary gestures to clarify intentions (e.g., gaze, pointing) ([54], [55]); and the greatest possible visibility of the robot motion ([56], [57]). We agree that these variables reported may have played a role in the scores. For example, the visibility of the robot motion may have been reduced by the video's recording features; the activities which are not carried out in daily life (e.g., grasping a tool to assemble an object in an industrial-like environment, like in Video 2) may have been responsible for reduced scores to our questionnaire.

b) No significant difference was found between participants' scores in the interaction condition: Independently of the presence or absence of another agent, we believe that the absence of a significant effect of the interaction condition may suggest that watching a robot performing a sequence of goal-oriented actions elicits in a human observer an intention attribution that is influenced more by the robot's goal than by the context. This is supported by Gazzola et al. [23] who showed that the goal of action might be more important for the activation of the mirror neuron system, which plays a role more in the understanding of the actions than the completion modality of the action [17].

For a better understanding of our results, with particular reference to section a) and section b), we believed that greater focus should be given to the similarities and differences characterizing the task's videos and to what we mean by "context." Our idea was to create a measure that could be used in different scenarios and not be limited to assessing the legibility of a robot's behavior in the restricted context of a specific task. For this reason, we wanted to explore how the readability of a robot's behavior may vary in different situations. Specifically, we decided to first investigate how an HRI scenario may affect this variable, compared to a scenario with a robot alone. In the selection phase, we chose six videos presenting a human–robot interaction scenario and six videos presenting a robot in a solitary setting. Of the first six videos, three of them presented a humanoid robot, while the other three presented a nonhumanoid robot. The same distinction was maintained for the other six videos. Despite this subdivision, the videos still have in common the presence of a robot performing a goal-oriented action (e.g., a behavior driven by an expectation that it is likely to bring about a desired outcome) [58]. Most of the videos also include a manipulation task (e.g., grasping an object, or folding laundry). This similarity may account for the absence of an interaction effect, meaning that despite the presence of a human–robot interaction or not, what makes a behavior readable is the comprehension of the goal itself. Therefore, physical features such as the presence of an object and whether it is visible, may have a leading role in the capacity to infer on robot's actions in different scenarios as features that are closely related to the target.

c) Participants performed better in videos showing humanoid robots than nonhumanoid robots: Our results also showed that people find it easier to understand the behavior of a human-like robot rather than a nonhumanoid robot. This is in line with the existing literature that highlights the role of human-like appearance in facilitating mind attribution ([18], [21], [22]). Our finding can also be considered in line with the "complementary gesture" assumption, according to which the use of complementary motions made by the robot could improve legibility (e.g., during an arm motion looking at the object) [59]. This might be the case in humanoid robot videos shown in our measure's validation, where the robot's head faced and followed the object of interest in goal-oriented actions. This robot's behavior is a forethought cue that makes the robot's behavior more readable. For example, in Video 7, the robot turns its head toward origami before grasping it.

d) Participants performed better in videos without explanations than with explanations: Contrary to our expectations, symbolic explanations seem to worsen the scores (i.e., making the robot's actions less readable). However, we observed that in videos featuring a nonhumanoid robot, explanations may have a facilitating effect, although not significant. Indeed, participants scored higher on videos with explanations and a nonhumanoid robot than on videos with explanations and a humanoid robot. The legibility and predictability of the robot's actions increase in the scenario with a nonhumanoid robot providing explanations. We believe that, in this case, since participants were not able to attribute a mental model similar to humans, people were able to predict the robot's behaviors with the help of verbal descriptions. In the case of humanoid robots, textual explanations were a distracting or annoying factor for the participants, which is in line with recent studies exploring the effect of the timing of explanations on people's perception of a robot and its behavior. Indeed, Han et al. [42] found that approximately half of the participants who watched videos of robots performing action-oriented movements wanted the robot to explain unexpected events as they happened, and only a small percentage of participants preferred the explanations from the robot before these happened. Another study also found that providing explanations of a robot's behavior before its execution rather than afterward makes the robot's behavior perceived as less desirable [60].

e) Participants performed better in the videos presenting short sequences rather than long and medium-long sequences: We hypothesized that participants were able to attribute intentions to the robot's action better when shown in short sequences rather than in medium or long sequences because sequencing tasks with a larger number of video clips may require a higher cognitive workload.

A. How to Use the HRIVST in Your Study

Our tool can be used by our peers for assessing people's ability to understand correctly a robot's behavior in different contexts in comparative design studies with different conditions (e.g., with or without explanations, different types of explanations, etc.). First, while creating the videos it is fundamental to consider our results which imply that

1) the surroundings and contextual environment need to be minimalized or controlled to keep a person's focus of attention active;

2) short interactions are considered more clear, hence if the goal is to compare different interactions, the length of the videos should be the same;
3) people are able to read more clearly behaviors of humanoid robots; and
4) the timing of the explanations is relevant, hence whether to provide them during the action, before or after should be carefully considered also considering the type of the provided explanation.

Then, the videos should be split into sequences of elementary actions. For example, if a robot is picking and placing an object in the video, this should be at least divided into three new video clips: 1) picking the object; 2) moving to the new position; and 3) placing the object. The video clips created are to be shown in a random order to the participants. Then, participants are asked to order the video clips according to the logical sequence of the actions. Finally, participants' responses should be collected to the questions as previously defined, and that identify the robot's intention, the person's expectation of the next robot's behavior, and the person's confidence in attributing the intention to the robot.

VII. Limitations and Future Work

Our current work presents some limitations that will guide future research. First, while some studies highlighted comparable results for assessing users' evaluation of robots' behavior using both video-based and in-person interaction studies [61], other studies suggest that people do not necessarily respond in the same way to robots in videos and physical scenarios [62]. Specifically, video scenarios presenting an interaction between a robot and a human were more effective in increasing users' perceptions of a robot's ease of use and adaptability (i.e., how well the robot completed its tasks) compared to a live interaction scenario, probably due to the fluency with which the video presented the interaction [62].

For a more precise evaluation of the effectiveness of explanations, we will explore different modalities for providing explanations considering different information provided as well as different ways to provide them (oral or written).

Finally, we intend to assess whether the individuals' characteristics (e.g., age) influence the legibility of the robot's behavior. Thus, we will focus on the extension of a diverse sample of participants to provide useful insights for enhancing human–robot interaction in different contexts.

VIII. Conclusion

In this work, we were interested in developing evaluation metrics for the legibility of the robot's behavior starting from the assessment of the ability of humans to attribute intentions to robots in different contexts. To these extents, we asked healthy subjects to evaluate the intentions of humanoid and nonhumanoid robots performing goal-oriented actions in videos with or without labels describing such actions. We evaluated their perceptions and the robot's legibility and predictability by a new measurement, HRIVST, here presented.

Our results show that HRIVST is a reliable tool for assessing the legibility of a robot's behavior by evaluating the human ability to correctly interpret its intentions. Our results also showed that people find the behaviors of a human-like robot more comprehensible than those of a nonhumanoid robot. This is also in line with the existing literature that underlies the role of human-like appearance in facilitating mind attribution. Moreover, our results have shed some light on the role of symbolic explanations in enhancing people's understanding of a robot's intentions, as assessed through video scenarios. Our results suggest a facilitating, although not significant, effect on the legibility of the intentions when the video included nonhumanoid robotic behaviors.

The finding of this study will be used to investigate further both the effects of individual characteristics and the role of subjective perception of the effectiveness of explanations in determining the legibility of a robot's behaviors.

Declarations

Conflict of Interest: The authors declare that they have no conflict of interest.

Ethics approval: The study was approved by the Ethical Committee of Psychological Research of the University of Naples Federico II (n. 13/2022).

Consent to participate: Participants gave written informed consent prior to participation.

References

[1] M. Szollosy, "Shifting the goalposts: Reconceptualizing robots, AI, and humans," in *Minding the Future*. Berlin, Germany: Springer, 2021, pp. 219–242.
[2] J. Scholtz, "Theory and evaluation of human robot interactions," in *Proc. IEEE 36th Annu. Hawaii Int. Conf. Syst. Sci.*, 2003.
[3] M. Matarese, A. Sciutti, F. Rea, and S. Rossi, "Toward robots' behavioral transparency of temporal difference reinforcement learning with a human teacher," *IEEE Trans. Human-Mach. Syst.*, vol. 51, no. 6, pp. 578–589, Dec. 2021.
[4] A. Rossi, M. M. Scheunemann, G. L'Arco, and S. Rossi, "Evaluation of a humanoid robot's emotional gestures for transparent interaction," in *Social Robotics*. Berlin, Germany: Springer, 2021, pp. 397–407.
[5] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *Proc. IEEE 5th Int. Conf. Data Sci. Adv. Analytics*, 2018, pp. 80–89.
[6] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surveys*, vol. 51, no. 5, pp. 1–42, Aug. 2018.
[7] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling, "Explainable agents and robots: Results from a systematic literature review," in *Proc. 18th Int. Conf. Auton. Agents Multiagent Syst.*, 2019, pp. 1078–1088.
[8] D. Premack and G. Woodruff, "Does the chimpanzee have a theory of mind?," *Behav. Brain Sci.*, vol. 1, no. 4, pp. 515–526, 1978.
[9] C. D. Frith and U. Frith, "Interacting minds–A biological basis," *Science*, vol. 286, no. 5445, pp. 1692–1695, 1999.
[10] S. H. Huang, D. Held, P. Abbeel, and A. D. Dragan, "Enabling robots to communicate their objectives," *Auton. Robots*, vol. 43, no. 2, pp. 309–326, 2019.
[11] S.-L. Lee, I. Y.-M. Lau, S. Kiesler, and C.-Y. Chiu, "Human mental models of humanoid robots," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2005, pp. 2767–2772.
[12] A. D. Dragan, K. C. Lee, and S.S. Srinivasa, "Legibility and predictability of robot motion," in *Proc. ACM/IEEE 8th Int. Conf. Hum.-Robot Interact.*, 2013, pp. 301–308.
[13] T. Hellström and S. Bensch, "Understandable robots-what, why, and how," *Paladyn, J. Behav. Robot.*, vol. 9, no. 1, pp. 110–123, 2018.
[14] L. Takayama, D. Dooley, and W. Ju, "Expressing thought: Improving robot readability with animation principles," in *Proc. 6th Int. Conf. Hum.-Robot Interact.*, 2011, pp. 69–76.

[15] C. Lichtenthäler and A. Kirsch, "Legibility of robot behavior: A literature review," Apr. 2016. [Online]. Available: https://hal.science/hal-01306977
[16] M. Brüne, "Theory of mind and the role of IQ in chronic disorganized schizophrenia," *Schizophrenia Res.*, vol. 60, no. 1, pp. 57–64, 2003.
[17] G. Rizzolatti and L. Craighero, "The mirror-neuron system," *Annu. Rev. Neurosci.*, vol. 27, pp. 169–192, 2004.
[18] M. C. Martini, C. A. Gonzalez, and E. Wiese, "Seeing minds in others–can agents with robotic appearance have human-like preferences?," *PLoS One*, vol. 11, no. 1, 2016, Art. no. e0146310.
[19] S. Thellman, M. de Graaf, and T. Ziemke, "Mental state attribution to robots: A systematic review of conceptions, methods, and findings," *ACM Trans. Hum.-Robot Interact.*, vol. 11, no. 4, pp. 1–51, 2022.
[20] S. Marchesi, D. Ghiglino, F. Ciardo, J. Perez-Osorio, E. Baykara, and A. Wykowska, "Do we adopt the intentional stance toward humanoid robots?," *Front. Psychol.*, vol. 10, 2019, Art. no. 450.
[21] S. Kiesler, A. Powers, S. R. Fussell, and C. Torrey, "Anthropomorphic interactions with a robot and robot–like agent," *Social Cogn.*, vol. 26, no. 2, pp. 169–181, 2008.
[22] C. E. Looser and T. Wheatley, "The tipping point of animacy: How, when, and where we perceive life in a face," *Psychol. Sci.*, vol. 21, no. 12, pp. 1854–1862, 2010.
[23] V. Gazzola, G. Rizzolatti, B. Wicker, and C. Keysers, "The anthropomorphic brain: The mirror neuron system responds to human and robotic actions," *Neuroimage*, vol. 35, no. 4, pp. 1674–1684, 2007.
[24] F. Hegel, S. Krach, T. Kircher, B. Wrede, and G. Sagerer, "Theory of mind (tom) on robots: A functional neuroimaging study," in *Proc. ACM/IEEE 3rd Int. Conf. Hum.-Robot Interact.*, 2008, pp. 335–342.
[25] F. Castelli, F. Happé, U. Frith, and C. Frith, "Movement and mind: A functional imaging study of perception and interpretation of complex intentional movement patterns," in *Social Neuroscience*. London, U.K.: Psychology Press, 2013, pp. 155–169.
[26] C. K. Morewedge, J. Preston, and D. M. Wegner, "Timescale bias in the attribution of mind," *J. Pers. Social Psychol.*, vol. 93, no. 1, pp. 1–11, 2007.
[27] J. Banks, "Theory of mind in social robots: Replication of five established human tests," *Int. J. Social Robot.*, vol. 12, no. 2, pp. 403–414, 2020.
[28] S. Thellman, A. Silvervarg, and T. Ziemke, "Some adults fail the false-belief task when the believer is a robot," in *Proc. Companion ACM/IEEE Int. Conf. Hum.-Robot Interact.*, 2020, pp. 479–481.
[29] E. Wiese, A. Wykowska, J. Zwickel, and H. J. Müller, "I see what you mean: How attentional selection is shaped by ascribing intentions to others," *PLOS ONE*, vol. 7, no. 9, pp. 1–7, Sep. 2012.
[30] N. Epley, A. Waytz, and J. T. Cacioppo, "On seeing human: A three-factor theory of anthropomorphism," *Psychol. Rev.*, vol. 114, no. 4, 2007, Art. no. 864.
[31] K. E. Powers, A. L. Worsham, J. B. Freeman, T. Wheatley, and T. F. Heatherton, "Social connection modulates perceptions of animacy," *Psychol. Sci.*, vol. 25, no. 10, pp. 1943–1948, 2014.
[32] A. Edwards, C. Edwards, D. Westerman, and P. R. Spence, "Initial expectations, interactions, and beyond with social robots," *Comput. Hum. Behav.*, vol. 90, pp. 308–314, 2019.
[33] A. Abubshait and E. Wiese, "You look human, but act like a machine: Agent appearance and behavior modulate different aspects of human–robot interaction," *Front. Psychol.*, vol. 8, 2017, Art. no. 1393.
[34] C. M. Eddy, "What do you have in mind? Measures to assess mental state reasoning in neuropsychiatric populations," *Front. Psychiatry*, 2019, Art. no. 425.
[35] R. Langdon, P. T. Michie, P. B. Ward, N. McConaghy, S. V. Catts, and M. Coltheart, "Defective self and/or other mentalising in schizophrenia: A cognitive neuropsychological approach," *Cogn. Neuropsychiatry*, vol. 2, no. 3, pp. 167–193, 1997.
[36] H. Wimmer and J. Perner, "Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception," *Cognition*, vol. 13, no. 1, pp. 103–128, 1983.
[37] L. Heavey, W. Phillips, S. Baron-Cohen, and M. Rutter, "The awkward moments test: A naturalistic measure of social understanding in autism," *J. Autism Dev. Disord.*, vol. 30, no. 3, pp. 225–236, 2000.
[38] I. Dziobek et al., "Introducing MASC: A movie for the assessment of social cognition," *J. Autism Devlop. Disord.*, vol. 36, no. 5, pp. 623–636, 2006.
[39] S. McDonald, S. Flanagan, J. Rollins, and J. Kinch, "Tasit: A new clinical tool for assessing social perception after traumatic brain injury," *Head Trauma Rehabil.*, vol. 18, no. 3, pp. 219–238, 2003.
[40] P. Haazebroek, S. Van Dantzig, and B. Hommel, "A computational model of perception and action for cognitive robotics," *Cogn. Process.*, vol. 12, no. 4, pp. 355–365, 2011.
[41] L. Kunze and M. Beetz, "Envisioning the qualitative effects of robot manipulation actions using simulation-based projections," *Artif. Intell.*, vol. 247, pp. 352–380, 2017.
[42] Z. Han, E. Phillips, and H. A. Yanco, "The need for verbal robot explanations and how people would like a robot to explain itself," *ACM Trans. Hum.-Robot Interact.*, vol. 10, no. 4, pp. 1–42, 2021.
[43] B. F. Malle, *How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction*. Cambridge, MA, USA: MIT Press, 2006.
[44] M. F. Folstein, S. E. Folstein, and P. R. McHugh, "'mini-mental state': A practical method for grading the cognitive state of patients for the clinician," *Psychiatr. Res.*, vol. 12, no. 3, pp. 189–198, 1975.
[45] E. A. Carlesimo, GA, "The mental deterioration battery: Normative data, diagnostic reliability and qualitative analyses of cognitive impairment," *Eur. Neurol.*, vol. 36, no. 6, pp. 378–384, 1996.
[46] G. Measso et al., "The mini-mental state examination: Normative study of an Italian random sample," *Devlop. Neuropsychol.*, vol. 9, no. 2, pp. 77–85, 1993.
[47] H. Spinnler and G. Tognoni, *Standardizzazione e Taratura Italiana di Test Neuropsicologici*. Masson Italia Periodici, 1987.
[48] J. C. Nunnally, *Psychometric Theory 3E*. New York: Tata McGraw-hill Education, 1994.
[49] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "A review of human activity recognition methods," *Front. Robot. AI*, vol. 2, 2015, Art. no. 28.
[50] A. Gupta, A. Kembhavi, and L. S. Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 10, pp. 1775–1789, Oct. 2009.
[51] M. Beetz et al., "Generality and legibility in mobile manipulation: Learning skills for routine tasks," *Auton. Robots*, vol. 28, pp. 21–44, 2010.
[52] J. Guzzi, A. Giusti, L. M. Gambardella, G. Theraulaz, and G. A. Di Caro, "Human-friendly robot navigation in dynamic environments," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2013, pp. 423–430.
[53] D. Bortot, M. Born, and K. Bengler, "Directly or on detours? how should industrial robots approximate humans?," in *Proc. ACM/IEEE 8th Int. Conf. Hum.-Robot Interact.*, 2013, pp. 89–90.
[54] C. L. Nehaniv, K. Dautenhahn, J. Kubacki, M. Haegele, C. Parlitz, and R. Alami, "A methodological approach relating the classification of gesture to identification of human intent in the context of human–robot interaction," in *Proc. IEEE Int. Workshop Robot Hum. Interact. Commun.*, 2005, pp. 371–377.
[55] P. Basili et al., "Inferring the goal of an approaching agent: A human-robot study," in *Proc. IEEE 21st Int. Workshop Robot Hum. Interact. Commun.*, 2012, pp. 527–532.
[56] F. Dehais, E. A. Sisbot, R. Alami, and M. Causse, "Physiological and subjective evaluation of a human–robot object hand-over task," *Appl. Ergonom.*, vol. 42, no. 6, pp. 785–791, 2011.
[57] E. A. Sisbot, A. Clodic, R. Alami, and M. Ransan, "Supervision and motion planning for a mobile manipulator interacting with humans," in *Proc. ACM/IEEE 3rd Int. Conf. Hum.-Robot Interact.*, 2008, pp. 327–334.
[58] A. Dickinson, "Actions and habits: The development of behavioural autonomy," *Philos. Trans. Roy. Soc. London. B., Biol. Sci.*, vol. 308, no. 1135, pp. 67–78, 1985.
[59] E. A. Sisbot and R. Alami, "A human-aware manipulation planner," *IEEE Trans. Robot.*, vol. 28, no. 5, pp. 1045–1057, Oct. 2012.
[60] S. Stange and S. Kopp, "Explaining before or after acting? how the timing of self-explanations affects user perception of robot behavior," in *Social Robotics*. Berlin, Germany: Springer, 2021, pp. 142–153.
[61] S. Woods, M. Walters, K. L. Koay, and K. Dautenhahn, "Comparing human robot interaction scenarios using live and video based methods: Towards a novel methodological approach," in *Proc. IEEE 9th Int. Workshop Adv. Motion Control*, 2006, pp. 750–755.
[62] Q. Xu et al., "Effect of scenario media on human-robot interaction evaluation," in *Proc. ACM/IEEE Hum.-Robot Interact.*, 2012, pp. 275–276.

Open Access funding provided by 'Università degli Studi di Napoli "Federico II"' within the CRUI CARE Agreement