Human vs Humanoid.

A Behavioral Investigation of the Individual Tendency to Adopt the Intentional Stance

Serena Marchesi[†]

S4HRI

Istituto Italiano di Tecnologia

Genova, Italy

serena.marchesi@iit.it

Nicolas Spatola

S4HRI

Istituto Italiano di Tecnologia

Genova, Italy

nicolas.spatola@iit.it

Jairo Perez-Osorio

S4HRI

Istituto Italiano di Tecnologia

Genova, Italy

jairo.perez-osorio@iit.it

Agnieszka Wykowska

S4HRI

Istituto Italiano di Tecnologia

Genova, Italy

agnieszka.wykowska@iit.it

ABSTRACT

Humans interpret and predict behavior of others with reference to mental states or, in other words, by adopting the intentional stance. The present study investigated to what extent individuals adopt the intentional stance towards two agents (a humanoid robot and a human). We asked participants to judge whether two different descriptions fit the behaviors of the robot/human displayed in photographic scenarios. We measured acceptance/rejection rate of the descriptions (as an explicit measure) and response times in making the judgment (as an implicit measure). Our results show that at the explicit level, participants are more likely to use mentalistic descriptions for the human agent and mechanistic descriptions for the robot. Interestingly, at the implicit level, we found no difference in response times associated with the robotic agent. We argue that, at the implicit level, both stances are processed as “equally likely” to explain the behavior of a humanoid robot, while at the explicit level there is an asymmetry in the adopted stance. Furthermore, cluster analysis on participants’ individual differences in anthropomorphism likelihood revealed that people with a high tendency to anthropomorphize tend to accept faster the mentalistic description. This suggests that the

decisional process leading to adoption of one or the other stance to adopt is influenced by individual tendency to anthropomorphize non-human agents.

CCS CONCEPTS

• Human-centered computing → Human-Computer Interaction (HCI) → Empirical studies in HCI

KEYWORDS

Intentional Stance, mental states, Human-Robot Interaction, response times, anthropomorphism

ACM Reference format:

Serena Marchesi, Nicolas Spatola, Jairo Perez-Osorio and Agnieszka Wykowska. 2021. Human vs Humanoid. A behavioral investigation of the individual tendency to adopt the intentional stance. In *Proceedings of 2021 ACM /IEEE International Conference on Human-Robot Interaction HRI '21*, March 08/11, Boulder (CO), USA. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3434073.3444663>

1 Introduction

Humans predict others’ actions based on the assumption that the observed behavior can be explained with reference to internal mental states (i.e., desires, beliefs and goals). Given the covert nature of these internal states, we have learned to detect and interpret social signals in order to infer the mental states and predict upcoming actions [1]. Daniel Dennett [2] defined this strategy of predicting other agents’ behavior with reference to their mental states as the intentional stance. We adopt the intentional stance towards other humans quite effortlessly and

[†]SM is also with Department of Computer Science, University of Manchester, Manchester, UK

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
HRI '21, March 8–11, 2021, Boulder, CO, USA

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-8289-2/21/03...\$15.00

<https://doi.org/10.1145/3434073.3444663>

ubiquitously. Research already demonstrated that humans can sometimes adopt the intentional stance also towards artificial agents [3 for review], a phenomenon related to the process of anthropomorphism, i.e., attributing human properties to non-human agents. To what extent humans readily adopt the intentional stance toward robots and how to measure this phenomenon remain unclear, although some research has addressed this problem and tried to operationalize such philosophical concept [4; 5]. The present paper addresses these issues by investigating whether humans would adopt the intentional stance in order to predict and explain the actions of humans and robots, at the explicit level (acceptance/rejection of a description that uses mechanistic/mentalistic vocabulary), and at the implicit level (participants' response time in the task). In addition, this paper investigates the relation between likelihood of anthropomorphism and intention attribution and attempt to delineate phenotypes of the tendency towards adopting intentional stance.

1.1 The Intentional Stance

According to Dennett [6, 2], in predicting and explaining behavior of other systems, humans employ the strategy that is most efficient. For example, we expect a kite to fly, as it is designed to do so. If one of its surfaces is broken, we expect that the kite would not rise high to the sky. Predicting behavior of a system using the designed function of the object is what Dennett defined as adopting the design stance. Alternatively, when it comes to other humans, the most efficient stance is the intentional stance, i.e., predicting behaviors based on inferred mental states. It has been extensively demonstrated that healthy adults spontaneously adopt the intentional stance towards other humans [for a review see 3]. Interestingly, humans also have the tendency to adopt intentional stance to non-human objects [7; 8; 9; 10]. Nevertheless, it is yet to be understood whether humans employ a similar social cognitive mechanism during interaction

with robots, and what conditions determine whether humans would choose one over the other stance.

1.2 Adopting the intentional stance towards robots and the Intentional Stance Test (IST)

In recent years, multiple studies have attempted to operationalize the adoption of intentional stance towards humanoid robots [see 3, 11 for review]. For example, Thellman et al., [4] evaluated whether people would use the same causal models to explain the behavior of a human and a humanoid robot. The authors reported that participants used (to a certain degree) similar intentional explanations behavior for both agents. They also showed that observers were less confident assigning intentional explanations to the robot. In a similar fashion, Marchesi and colleagues [5] proposed a test – the InStance Test (IST) – to evaluate whether people employ mentalistic or mechanistic explanations to describe the behavior of a humanoid robot. IST consisted of pictures (34 scenarios) showing the iCub robot [12] performing some daily-life actions, each one with a sequence of events depicted in three images (See figure 1). Participants were asked to select the description (mentalistic or mechanistic) that, according to them, better explained the behavior the robot. Authors reported that although participants had a bias toward the mechanistic explanations, in some circumstances, they opted for mentalistic explanations. Importantly, the authors suggest that adopting the intentional stance towards artificial agents does not imply that observers assume that those agents have indeed mental states that govern their behavior [see also 13]. Instead, people might predict artificial agents' behavior “as if” they would have mental states. Therefore, people might switch between adopting the intentional and design stances when they try to explain the behavior of an (artificial) agent. Interestingly, the results of Marchesi and colleagues [5] study revealed a binomial distribution on the responses, as participants tended to

Figure 1 Examples of items depicting either the iCub robot (Panel A, top) or the human character, Sara (Panel B, bottom) as protagonists of the scenarios.

“polarize” in the preferences for either mentalistic or mechanistic explanations of the robot actions. Furthermore, the human tendency to ascribe human-like characteristics, such as intentionality or emotions, to inanimate entities (known as anthropomorphism, [14, 15]), might also contribute to the attribution of human mental capabilities to robots [16, 17, 18]. Arguably, the attribution of human-like characteristics activates the general knowledge we have about humans [14; 15; 19] and the activation of neural representation related to the perception of other humans [20; 21; 22]. For example, factors like appearance [23], social gaze [24; 25], expectations and individual differences [26; 27], variability of behavior [28], and perception of human likeness [29], might also impact the level of attributed anthropomorphism and therefore help to induce the adoption of the intentional stance towards a humanoid robot. Finally, the adoption of intentional stance and anthropomorphic attributions may be influenced by individual characteristics (e.g., dispositional, developmental and cultural) and the context in which individuals is observing an agent [16]. In sum, it is pivotal to investigate the individual tendency of humans to adopt the intentional stance, to take into account the “intentional stance tendency phenotype” of each individual.

1.3 Task adaptation

In the original IST of Marchesi et al. [5], participants were asked to provide their subjective rating (on a continuous scale) regarding the descriptions of the scenarios they viewed. To have a further insights into the processes leading to the adoption of the intentional stance, we adapted Marchesi et al. task to a design with a two-alternative-forced-choice task (2AFC) [30]. Such adaptation grants for collecting participants’ response times (RTs), a classical mental chronometry measure of cognitive processes used in experimental psychology and HRI [31; 32; 33; 34; 35; 36; 60; 17]. Therefore, our adaptation makes it possible to evaluate the adoption of the intentional stance at a more implicit level. Explicit attitudes operate on a conscious level and are generally measured through explicit self-reports (e.g. questionnaires), while implicit attitudes rely on unconscious and automatic processes, and are typically assessed via implicit measures (e.g. response time paradigms, implicit association test) [37]. Research suggests that implicit attitudes might constitute better predictors of future intentions and behaviors [38], and thus be more representative of real attitudes than explicit declarations. Implicit measures have also proved to be well equipped to predict the behavioural consequences of individuals’ implicit representations [39; 38].

1.4 Aims and hypotheses

The aims of the present study were to (i) test whether participants’ response times would differ while processing the two stances and (ii) replicate and validate the results of Marchesi et al. [5]. Furthermore, we included a control condition in which a human actor was presented in same actions as the robot agent, allowing us to compare participants’ behavior between the two

agents, thereby further validating the IST. We hypothesized that participants would (1) accept more often (explicit level) and have faster response times (implicit level) for mentalistic statements in human scenarios. We expected that participants would be less likely to accept (explicit level) and have slower response times (implicit level) for mentalistic statements in the scenarios with iCub. The secondary aim of the study was to (2) investigate the inter-individual differences in the tendency to anthropomorphize by measuring how the likelihood to anthropomorphize robots may predict the likelihood of adopting the intentional stance towards a humanoid robot.

2 MATERIALS AND METHODS

Forty-one participants were recruited for the experiment and received a monetary compensation of 15€. One participant was excluded from the analysis due to a technical problem in one of the blocks, leading to a final sample of $N = 40$ (mean age: 24.77; SD: 4.01; 25 women). All participants reported no history of psychiatric or neurological diseases and signed the informed consent before the beginning of the experimental session. The study was approved by the local Ethical Committee (Comitato Etico Regione Liguria) and was conducted in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki). All participants were naïve to the purposes of the study but were debriefed after the experiment.

2.1 Experimental design

The new version of the IST measured the ratio of acceptance/rejection of mentalistic/mechanistic statements and response times of participants’ choices. That is, two major modifications were made to the original IST. First, only one statement (mechanistic or mentalistic) was presented at a time for each scenario; second, instead of a slider, participants were asked to accept or reject each single description (mentalistic/mechanistic) associated with each scenario. This would result in a two-alternative forced-choice task. This design provides measures of participants’ performance on two levels: an implicit level (response times) and an explicit level (the acceptance/rejection rate). The test had two blocks, one with iCub was the main character (robot block) and another one in which a human (named Sara - the human block). Each trial consisted in a scenario taken from the original IST depicting the character performing an action and a sentence describing the agents’ behavior. The description presented below the scenarios was either mechanistic (“iCub/Sara was unbalanced for a moment”) or mentalistic (“iCub/Sara was trying to cheat by looking at opponent’s cards”). Each one of the original IST 34 scenarios from [5] was shown twice: once with the mechanistic description and once with the mentalistic description. This led to 68 items per block (robot and human). The items for the robot block were taken from the InSTance Test (IST, [5]); while items for the human consisted of photos of equivalent scenarios but depicting a human character (Sara) instead of the iCub (see Fig 1). Note that the name of the

character (iCub or Sara) contains an equal number of letters, in order to avoid biases during the reading process. The descriptions were identical for both agents for each corresponding scenario. The presentation of the items within the block was randomized and the order of the human and robot blocks was counterbalanced across participants.

2.2 Apparatus and procedure

The experiment was performed inside a dimly lit room. Items were presented on a 27" LCD screen (resolution: 1920 × 1080). Participants were seated approximately at 62 cm distance from the screen. After participants read and signed the consent form, they were instructed about the task. They were told that their task would be to judge whether the sentence presented below a scenario plausibly explained what was depicted in the scenario. Participants were asked to respond by pressing two keys on the numeric pad of the keyboard (8 for "yes" and 2 for "no"). Before each block, participants had a practice trial with the agent corresponding to the respective block. The trial procedure was the following: 1- a fixation dot appeared at the center of the screen for 500 ms; 2- the sequence of three images was presented at once at the center of the screen for 3000 ms. 3- the sentence appeared below the scenario with response options "yes" or "no" and remained until the keypress. Participants' response times were measured from the onset of the sentence to the keypress. Before the beginning of the human block, we asked participants to complete the Individual Differences in Anthropomorphism Questionnaire (IDAQ), developed by Waytz, Cacioppo and Epley [40]. The reason to present the IDAQ to the participants before the human block was two-folded: (1) to avoid any anthropomorphism bias towards the robot; (2) since observing a human may already induce anthropomorphism in the subsequent questionnaire, affecting the IDAQ score, a safe solution was to administer it before any exposure to the human agent.

2.3 The Individual Differences in Anthropomorphism Questionnaire (IDAQ)

In order to evaluate whether the individual differences in the general tendency towards anthropomorphism influences the attribution of mental states to artificial agents, we administered the Individual Differences in Anthropomorphism Questionnaire (IDAQ), by Waytz, Cacioppo & Epley [40]. The IDAQ investigates participants' stable and general individual differences in anthropomorphism with several items on various agents. The scale of the IDAQ items ranges from 0 to 10, with a Cronbach alpha $\geq .82$. Examples of the items are: "To what extent does the average insect have a mind of its own?", "To what extent do cows have intentions?" and "To what extent does the average fish have free will?". The authors argue that the mental state attributes used in the items (i.e. "have a mind of its own", "have intentions", and "have free will") have been proven to be reliable characteristics of human-uniqueness and higher order cognitive processes [41; 42; 43]. According to the authors, individual scores in IDAQ allow predicting three major factors

related to anthropomorphism that might have consequences in everyday life: 1- the moral care and concern towards other agents; 2- the trust and responsibility attributed to other agents and 3- the extent to which an agent is considered as root of social influence on the self. Therefore, the IDAQ can be a useful tool to investigate at the individual level, how the likelihood of anthropomorphism can relate to the adoption of the intentional stance strategy towards a robotic agent.

3 RESULTS

Statistical analysis were conducted in R Studio (version 4.0.2, [44]) using the package lme4 [45].

3.1 Acceptance/Rejection of mentalistic description

To investigate the probability of acceptance or rejection of a description, we performed a Generalized Linear Mixed Model. We considered the choice (acceptance/rejection) as variable of interest, while the type of description (mechanistic vs. mentalistic) and the agent (human vs. robot) as fixed within-participants factors. Participant and scenario were considered as random factors. Results showed a significant interaction effect between type of description*agent: $b = -2.30$, $z(5433) = -19.03$, $p = <.001$, $CI_{95\%} [-2.53; -2.06]$ (comparison with null model: $\chi^2(1) = 381.03$, $p = <.001$). As visible in Figure 2, participants showed an inverse pattern of mentalistic/mechanistic description acceptance for the two agents. When asked explicitly, they accepted more often the mentalistic description for a human while they accepted more often the mechanistic description. We further investigated the contrast between mechanistic and mentalistic description with planned post-hoc pairwise comparisons (Tukey's HSD correction for multiple comparisons): human agent: $b = -1.70$, $p = <.001$; robot agent: $b = .60$, $p = <.001$.

3.2 Response Times results

The Linear Integrated Speed-Accuracy Score (LISAS) approach

Trials with RTs slower or faster than 3 standard deviations per condition (mechanistic vs. mentalistic; human vs. robot) for each participant were considered outliers and then removed from analyses. This resulted in the exclusion of 1.27% (69) of the trials (this filtering procedure has the advantage of taking out extreme values without affecting the data of one specific condition or of one participant in particular). To investigate participants' response times during the task, we analyzed the relation between response times, agent presented and descriptions, correcting participants' response time using the Linear Integrated Speed-Accuracy Score (LISAS, [46]).

Figure 2 interaction effect in GLMM between description*agent. The y-axis reports the predicted probability of acceptance/rejection on a binomial scale: positive numbers represents the probability of acceptance (yes) and the negative numbers represent the probability of rejection (no). * = $p < .05$, ** = $p < .01$, *** = $p < .001$

This transformation allows considering a linear relationship between response times (RT) and the proportion of decision (PD) toward a response 0 and an alternative response 1. This makes it possible to evaluate participants' response bias including both the decision type and the decision speed. The main advantage of this analysis method (compared to standard RT analysis) consists in including and integrating the type of choice and response time for each condition (and their respective variance) in a single score. We considered the proportion of choice (acceptance/rejection) associated with the proportion of mentalistic description choices (i.e., 1) contrasted to the mechanistic description choices (i.e., 0).

$$\text{LISAS} = \text{RT}_x + \text{SRT}/\text{SPD} \times \text{PD}_x \times (1)$$

Where RT_x is the mean response time in condition x , PD_x is the proportion of choice (1 vs 0) in condition x and SRT and SPD is the overall standard deviation for response times and proportion of choices, respectively. Two additional participants that exclusively chose mechanistic or mentalistic responses in all trials in at least one condition were excluded from the analysis since their data are to be considered not reliable. The remaining sample size was equal to 38 ($N = 38$). We conducted a linear mixed effect model with agent and description as fixed factors and RTs -LISAS corrected as a variable of interest. Participant was considered as random factor. We found an

interaction effect between agent and description ($b = -2467.2$, $t(38) = -3.94$, $p < .001$, $\text{CI}_{95\%} [-3688.84; -1245.60]$) (Fig.3), showing that participants were faster in accepting a mentalistic description for the human agent, relative to the robotic agent. Contrasts with Bonferroni correction showed that participants were faster in choosing the mentalistic descriptions in the human compared to robot trials, $F(1, 37) = 10.34$; $p = .003$; $\text{CI}_{95\%} [495.16; 2182.41]$. Conversely, they were faster choosing the mechanistic descriptions in robot compared to human trials, $F(1, 37) = 6.93$; $p = .012$; $\text{CI}_{95\%} [259.66; 1997.22]$. The difference between response times between mentalistic and mechanistic was significant only in human trials, $F(1, 37) = 74.46$; $p < .001$; $\text{CI}_{95\%} [1998.01; 3224.26]$, with a faster responses for mentalistic compared to mechanistic descriptions.

Figure 3 Interaction effect on LISAS scores agent*description.* = $p < .05$, ** = $p < .01$, *** = $p < .001$.

Clustering approach on the individual tendency to accept a mentalistic description for a robot and the IDAQ score

To evaluate whether we could categorize participants in order to predict their tendency for anthropomorphism, we first processed a two-step clustering using the mentalistic (vs. mechanistic) bias of participants measured by LISAS score [46]. Given the way LISAS score are computed, they contain information about the relationship between the response times and the proportion of choice (acceptance/rejection) associated with the mentalistic and

mechanistic descriptions. The result is a score that represents the mentalistic bias of each participant. We preferred the 2 step-clustering approach (compared to correlation) due to the high inter-individual variability in RTs range, that biases the correlation results with the normative IDAQ scale. As we used LISAS score and only have one data point for each participant for each condition (4 in total), a normalization would be highly sensitive to this inter-individual variability. We thus preferred the cluster approach that provides a common structure to all participants. The clustering proposed a solution with a 2 clusters matrix with a 1.71 ratio sizes and a cluster quality = 0.7 that measure the cohesion and separation of clusters (good fit) (for similar procedure see [47]). According to cluster silhouette and cluster comparison, analyses argue for a high tendency vs. low tendency group in mentalistic attributions toward robot. Indeed, participants in the first cluster chose the mentalistic description for the robot more rapidly compared to participants in the second cluster, $F(1, 37) = 56.94$, $p < .001$, $CI95\% [-4105.78; -1239.28]$. Interestingly, the clusters also predicted participants' tendency to anthropomorphize (IDAQ scores). Participants in the cluster 1 scored higher on the IDAQ than participants in the cluster 2, $F(1, 37) = 4.28$, $p = .046$, $CI95\% [.29, 30.34]$.

4 DISCUSSION

The aims of the present paper were to: (i) test whether participants' response times would differ while deciding to adopt one the two stances towards a human or the robot and (ii) replicate and validate the results of Marchesi et al. [5]. In addition, we aimed at examining the relationship between likelihood of adopting the intentional stance and individual tendency to anthropomorphize. In line with our hypothesis, at the explicit level (Acceptance/Rejection of mentalistic description), we found that mechanistic descriptions were attributed to the robot to a higher extent than mentalistic descriptions. The inverse pattern was found for the human, where participants were more likely to attribute a mentalistic description to the depicted human. Indeed these results replicate the results of Marchesi et al. [5] with a different design and measure. In Marchesi et al. [5], the authors reported an average IST score of 40.73 on a scale from 0 to 100 (where 0 would be totally mechanistic and 100 totally mentalistic), showing a bias towards the mechanistic description for the robot. The present study made it possible to use objective measure such as response times as a probe for adopting the intentional stance. The results showed that participants' response times were faster in mentalistic descriptions for a human compared to a robot, but also compared to mechanistic descriptions. Furthermore, participants were faster in responding in mechanistic descriptions for the robot, as compared to the human. Interestingly, when looking only at the robotic agent condition, no significant difference in the response times between mechanistic and mentalistic descriptions was observed. This is in contrast to the human condition, where the mentalistic descriptions were evaluated faster than mechanistic

descriptions. This pattern suggests that the intentional stance might be the default strategy to explain other humans' behaviors both at the explicit and implicit level. When it comes to a robotic agent, participants might not find any of the descriptions as unlikely as the mechanistic descriptions for the human. Meaning that both stances are compatible with participants' mindset, at least at the implicit level. However, at the explicit level, participants are more likely to accept mechanistic descriptions of the robot actions. This might reflect that both the intentional and design stance are strategies that could be adopted to explain an artificial agent's behavior. Some authors assume that intentional stance would be the default stance in general (relying on the social cognition system) [48; 49]. Indeed, neural areas (e.g., the superior temporal sulcus, lateral fusiform gyrus, medial prefrontal cortex, posterior cingulate, insula and amygdala,) associated with mentalizing (*i.e.* the attribution of mental states to others), belong to the default mode network [50], suggesting that this key social cognition mechanism is the default mode of functioning of the brain [50; 51; 52]. However, although a default system, it can be controllable [53], this would imply that at the earlier (implicit) stage of processing, individuals may adopt the intentional stance towards a robot but upon deliberate reflection, they would quickly identify the robot as an artifact, and they would reject the mentalistic descriptions. Results from cluster analyses revealed that participants differed in the tendency to adopt the intentional stance, showing the presence of individual differences in explaining a humanoid robot's behavior with a mentalistic vocabulary [13]. Moreover, the results showed that those participants that were faster in choosing a mentalistic description for the robot, also scored higher in the IDAQ questionnaire, as compared to those who were slower in selecting mentalistic descriptions. This is in line with previous literature [54] and shows that a high level of anthropomorphism is associated with a high tendency to adopt the intentional stance. Moreover, recent literature reports that participants' tendency to adopt the intentional or design stance, could be influenced by individual differences [55, 27, 56], the need for cognition [57, 58] or the expectations about robots [26] but also the embedded cultural values and other factors [16]. In line with these previous studies [16], our results argue that it might be useful to delineate personality phenotypes to describe clusters of individual characteristics, such as the tendency to anthropomorphize, since they might be predictive of future interactions in HRI.

5 LIMITATION AND FUTURE WORK

The present adaptation of Marchesi et al. [5] material, despite using RT, remains a choice task between two options about a complex semantic sentence. According to the RT that are longer (overall LISAS mean: 6749.77)) than usual implicit test using implicit association (RT mean = 2500-3000) [37; 59], we cannot rule out that the decision could suffer from biases inherent to explicit reasoning engaged by the participants. Moreover, although we cannot exclude that varying confederates across

different items may affect participants' performance, we highlight that in [5], the authors showed that humans included in scenarios do not cause substantially different scores relative to non-human scenarios. Future, tailored, work should investigate these aspects.

6 CONCLUSIONS

Our findings indicate that it is possible to design a test probing adoption of intentional stance with objective measures, such as response times. Furthermore, we show that the individual bias in the adoption of the intentional stance towards humanoid robots should be investigated both on the explicit and implicit level, as those levels are not necessarily always in accordance [see also 27]. Additionally, the evaluation of individual tendencies to adopt the intentional stance and to anthropomorphize allowed exploring more in depth the connection between individual preferences for intentional stance towards artificial agents and anthropomorphism.

The study dataset is available in the OSF repository at <https://osf.io/65bn2/>

ACKNOWLEDGMENTS

This work has received support from the European Research Council under the European Union's Horizon 2020 research and innovation program, ERC Starting grant, G.A. number: ERC-2016-StG-715058, awarded to Agnieszka Wykowska. The content of this paper is the sole responsibility of the authors. The European Commission or its services cannot be held responsible for any use that may be made of the information it contains.

REFERENCES

- [1] Simon Baron-Cohen, Howard A. Ring, Sally Wheelwright, Edward T. Bullmore, Mick J. Brammer, Andrew Simmons, & Steve C. R. Williams, 1999. Social intelligence in the normal and autistic brain: an fMRI study. *European journal of neuroscience*, 11(6), 1891-1898. doi.org/10.1046/j.1460-9568.1999.00621.x
- [2] Daniel C. Dennett, 1989. *The intentional stance*. MIT press
- [3] Jairo Perez-Osorio & Agnieszka Wykowska, 2020. Adopting the intentional stance toward natural and artificial agents. *Philosophical Psychology*, 33(3), 369-395. doi.org/10.1080/09515089.2019.1688778.
- [4] Sam Thellman, Annika Silvervarg & Tom Ziemke, 2017. Folk-psychological interpretation of human vs. humanoid robot behavior: Exploring the intentional stance toward robots. *Frontiers in psychology*, 8, 1962. doi.org/10.3389/fpsyg.2017.01962
- [5] Serena Marchesi, Davide Ghiglini, Francesca Ciardo, Jairo Perez-Osorio, Ebru Baykara, & Agnieszka Wykowska, 2019. Do we adopt the intentional stance toward humanoid robots?. *Frontiers in psychology*, 10, 450. doi.org/10.3389/fpsyg.2019.00450.
- [6] Daniel C. Dennett, 1971. Intentional systems. *The Journal of Philosophy*, 87-106.
- [7] Fritz Heider & Marianne Simmel, 1944. An experimental study of apparent behavior. *The American journal of psychology*, 57(2), 243-259. doi.org/10.2307/1416950.
- [8] Jan Zwickel, 2009. Agency attribution and visuospatial perspective taking. *Psychonomic Bulletin & Review*, 16(6), 1089- 1093. [Doi.org/10.3758/PBR.16.6.1089](https://doi.org/10.3758/PBR.16.6.1089).
- [9] Francesca Happé & Uta Frith, 1995. Theory of mind in autism. In *Learning and cognition in autism* (pp. 177-197). Springer, Boston, MA. doi.org/10.1007/978-1-4899-1286-2_10.
- [10] Ahamad M. Abu-Akel, Ian A. Apperly, Stephen J. Wood & Peter C. Hansen, 2020. Re-imaging the intentional stance. *Proceedings of the Royal Society B*, 287(1925), 20200244. doi.org/10.1098/rspb.2020.0244.
- [11] Elef Schellen & Agnieszka Wykowska, 2019. Intentional mindset toward robots—open questions and methodological challenges. *Frontiers in Robotics and AI*, 5, 139. doi.org/10.3389/frobt.2018.00139.
- [12] Giorgio Metta, Lorenzo Natale, Francesco Nori, Giulio Sandini, David Vernon, Luciano Fadiga, Claes von Hofsten, Kerstin Rosander, Manuel Lopes, José Santos-Victor, Alexandre Bernardino & Luis Montesano, 2010. The iCub humanoid robot: An open-systems platform for research in cognitive development. *Neural networks*, 23(8-9), 1125-1134. [Doi.org/10.1016/j.neunet.2010.08.010](https://doi.org/10.1016/j.neunet.2010.08.010).
- [13] Clifford Nass & Youngme Moon, 2000. Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1), 81-103. doi.org/10.1111/0022-4537.00153.
- [14] Gabriella Airenti, 2018. The development of anthropomorphism in interaction: Intersubjectivity, imagination, and theory of mind. *Frontiers in psychology*, 9, 2136. doi.org/10.3389/fpsyg.2018.02136.
- [15] Harriet Cullen, Ryota Kanai, Behador Bahrami & Geraint Rees, 2014. Individual differences in anthropomorphic attributions and human brain structure. *Social Cognitive and Affective Neuroscience*, 9(9), 1276-1280. doi.org/10.1093/scan/nst109.
- [16] Nicholas Epley, Adam Waytz & John T. Cacioppo, 2007. On seeing human: a three-factor theory of anthropomorphism. *Psychological review*, 114(4), 864. doi.org/10.1037/0033-295X.114.4.864.
- [17] Nicolas Spatola, Clément Belletier, Pierre Chausse, Maria Augustinova, Alice Normand, Vincent Barra, Ludovic Ferrand & Pascal Huguet, 2019. Improved cognitive control in presence of anthropomorphized robots. *International Journal of Social Robotics*, 11(3), 463-476. [Doi.org/10.1007/s12369-018-00511-w](https://doi.org/10.1007/s12369-018-00511-w).
- [18] Sara Kiesler, Aaron Powers, Susan R. Fussell & Cristen Torrey, 2008. Anthropomorphic interactions with a robot and robot-like agent. *Social Cognition*, 26(2), 169-181. doi.org/10.1521/soco.2008.26.2.169.
- [19] Gabriella Airenti, Marco Cruciani & Alessio Plebe, 2019. The cognitive underpinnings of anthropomorphism. *Frontiers in psychology*, 10, 1539. doi.org/10.3389/fpsyg.2019.01539.
- [20] Thierry Chaminade, Delphine Rosset, David Da Fonseca, Bruno Nazarian, Ewald Lutscher, Gordon Cheng & Christine Deruelle, 2012. How do we think machines think? An fMRI study of alleged competition with an artificial intelligence. *Frontiers in human neuroscience*, 6, 103. doi.org/10.3389/fnhum.2012.00103.
- [21] Helen L. Gallagher, Anthony I. Jack, Andreas Roepstorff & Christopger D. Frith, 2002. Imaging the intentional stance in a competitive game. *Neuroimage*, 16(3), 814-821. [Doi.org/10.1006/nimg.2002.1117](https://doi.org/10.1006/nimg.2002.1117).
- [22] Ceylan Özdem, Eva Wiese, Agnieszka Wykowska, Hermann Müller, Marcel Brass & Frank Van Overwalle, 2017. Believing androids—fMRI activation in the right temporo-parietal junction is modulated by ascribing intentions to non-human agents. *Social Neuroscience*, 12(5), 582- 593. doi.org/10.1080/17470919.2016.1207702.
- [23] Molly Martini, Christian A. Gonzalez & Eva Wiese, 2016. Seeing minds in others—Can agents with robotic appearance have human-like preferences?. *PloS one*, 11(1), e0146310. doi.org/10.1371/journal.pone.0146310.
- [24] Agnieszka Wykowska, Jasmin Kajopoulos, Miguel Obando-Leiton, Sushil Singh Chauhan, John-John Cabibihan & Gordon Cheng,

2015. Humans are well tuned to detecting agents among non-agents: examining the sensitivity of human perception to behavioral characteristics of intentional systems. *International Journal of Social Robotics*, 7(5), 767-781. doi.org/10.1007/s12369-015-0299-6.
- [25] Eva Wiese, Patrick Peter Weis & Daniel M. Lofaro, 2018. Embodied social robots trigger gaze following in real-time HRI. In *2018 15th IEEE International Conference on Ubiquitous Robots (UR)*, (June, 2018), pp. 477-482. Doi.org/10.1109/URAI.2018.8441825.
- [26] Jairo Perez-Osorio, Serena Marchesi, Davide Ghiglini, Melis Ince & Agnieszka Wykowska, 2019. More Than You Expect: Priors Influence on the Adoption of Intentional Stance Toward Humanoid Robots. In *International Conference on Social Robotics* (2019, November) pp. C1-C1. Springer, Cham.
- [27] Davide Ghiglini, Davide De Tommaso, Cesco Willemse, Serena Marchesi & Agnieszka Wykowska, 2020. Can I get your (robot) attention? Human sensitivity to subtle hints of human-likeness in a humanoid robot's behavior. (May, 2020). https://doi.org/10.31234/osf.io/kfy4g.
- [28] Serena Marchesi, Jairo Pérez-Osorio, Davide De Tommaso & Agnieszka Wykowska, 2020. Don't overthink: fast decision decision making combined with behavior variability perceived as more human-like. *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, Naples, Italy, (June, 2020), pp. 54-59, doi.org/10.1109/RO-MAN47096.2020.9223522.
- [29] Cesco Willemse & Agnieszka Wykowska, 2019. In natural interaction with embodied robots, we prefer it when they follow our gaze: a gaze-contingent mobile eyetracking study. *Philosophical Transactions of the Royal Society B*, 374(1771), 20180036. doi.org/10.1098/rstb.2018.0036.
- [30] Christoph Naefgen, Micheal Dambacher & Markus Janczyk, 2018. Why free choices take longer than forced choices: evidence from response threshold manipulations. *Psychological Research*, 82, 1039-1052. Doi.org/ 10.1007/s00426-017-0887-1
- [31] Fanciscus C. Donders, 1969/1868. On the speed of mental processes. *Acta psychologica*, 30, 412-431. doi.org/10.1016/0001-6918(69)90065-1
- [32] Arthur R. Jensen, 1987. Mental chronometry in the study of learning disabilities. *Mental Retardation & Learning Disability Bulletin*.
- [33] Duncan R. Luce, 1986. Response times: Their role in inferring elementary mental organization (No. 8). Oxford University Press on Demand.
- [34] Micheal I. Posner, 1978. *Chronometric explorations of mind*. Lawrence Erlbaum.
- [35] Anna Stenzel, Eriz Chinellato, Maria A. T. Bou, Ángel P. del Pobil, Markus Lappe & Roman Liepelt, 2012. When humanoid robots become human-like interaction partners: corepresentation of robotic actions. *Journal of Experimental Psychology: Human Perception and Performance*, 38(5), 1073. doi.org/10.1037/a0029493.
- [36] Kyveli Kompatsiari, Francesca Ciardo, Vadim Tikhonoff, Giorgio Metta & Agnieszka Wykowska, 2018. On the role of eye contact in gaze cueing. *Scientific reports*, 8(1), 1-10. doi.org/10.1038/s41598-018-36136-2.
- [37] Jan De Houwer, Sarah Teige-Mocigemba, Adrian Spruyt, & Agnes Moors, 2009. Implicit measures: A normative analysis and review. *Psychological bulletin*, 135(3), 347. doi.org/10.1037/a0014211.
- [38] Benedek Kurdi, Allison E. Seitchik, Jordan R. Axt, Timothy J. Carroll, Arpi Karapetyan, Neela Kaushik, Diana Tomezsko, Anthony G. Greenwald & Mahzarin R. Banaji, 2019. Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *American psychologist*, 74(5), 569. doi.org/10.1037/amp0000364.
- [39] Malte Frieze, Wilhem Hofmann & Michaela Wänk, 2008. When impulses take over: Moderated predictive validity of explicit and implicit attitude measures in predicting food choice and consumption behaviour. *British Journal of Social Psychology*, 47(3), 397-419. doi.org/10.1348/014466607X241540.
- [40] Adam Waytz, John Cacioppo, & Nicholas Epley, 2010. Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, 5(3), 219-232. doi.org/10.1177/1745691610369336.
- [41] Stéphanie Demoulin, Jacques.Philippe Leyens, Maria Paola Paladino, Ramon Rodriguez- Torres, Armando Rodriguez-Perez & John Dovidio, 2004. Dimensions of "uniquely" and "non-uniquely" human emotions. *Cognition and emotion*, 18(1), 71-96. doi.org/10.1080/02699930244000444.
- [42] Nick Haslam, Paul Bain, Lauren Douge, Max Lee & Brock Bastian, 2005. More human than you: Attributing humanness to self and others. *Journal of personality and social psychology*, 89(6), 937. doi.org/10.1037/0022-3514.89.6.937.
- [43] Megan Kozak, Abigail A. Marsh & Daniel M. Wegner, 2006. What do i think you're doing? Action identification and mind attribution. *Journal of personality and social psychology*, 90(4), 543. Doi.org/ 10.1037/0022-3514.90.4.543.
- [44] Team, R. C., 2013. R: A language and environment for statistical computing.
- [45] Douglas Bates, Martin Maechler, Ben Bolker, Steven Walker, Rune H. B. Christensen, Henrik Singmann, Bin Dai, Fabian Scheipl, Gabor Grothendieck & Green, P., 2018. Package 'lme4'. Version, 1, 17. https://CRAN.R-project.org/package=lme4
- [46] André Vandierendonck, 2016. A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. *Behavior Research Methods*, 49, 653- 673. Doi.org/10.3758/s13428-016-0721-5.
- [47] Nicolas Spatola & Karolina Urbanska, 2020. God-like robots: the semantic overlap between representation of divine and artificial entities. *AI & SOCIETY*, 35(2), 329-341. Doi.org/ 10.1007/s00146-019-00902-1.
- [48] Anthony I. Jack, Abigail J. Dawson & Megan E. Norr, 2013. Seeing human: Distinct and overlapping neural signatures associated with two forms of dehumanization. *NeuroImage*, 79, 313-328. Doi.org/ 10.1016/j.neuroimage.2013.04.109.
- [49] Leonard Schilbach, Simon B. Eickhoff, Anna Rotarska-Jagiela, Gereon R. Fink & Kai Vogeley, 2008. Minds at rest? Social cognition as the default mode of cognizing and its putative relationship to the "default system" of the brain. *Consciousness and Cognition*, 17(2), 457-467. doi.org/10.1016/j.concog.2008.03.013.
- [50] R. Nathan Spreng & Jessica R. Andrews-Hanna, 2015. The Default Network and Social Cognition. In *Brain Mapping: An Encyclopedic Reference*. doi.org/10.1016/B978-0-12-397025-1.00173-1.
- [51] Anthony Jack, Abigail J. Dawson, Katelyn L. Begany, Regina L. Leckie, Kevin Barry, Agneta H. Ciccio & Abraham Z. Snyder, (2013). fMRI reveals reciprocal inhibition between social and physical cognitive domains. *NeuroImage*, 66, 385-401. Doi.org/10.1016/j.neuroimage.2012.10.061.
- [52] R. Nathan Spreng, W. Dale Stevens, Joseph D. Viviano & Daniel L. Schacter, 2016. Attenuated anticorrelation between the default and dorsal attention networks with aging: evidence from task and rest. *Neurobiology of aging*, 45, 149-160. Doi.org/ 10.1016/j.neurobiolaging.2016.05.020.
- [53] Esmeralda G. Urquiza-Haas & Kurt Kotrschal, 2015. The mind behind anthropomorphic thinking: Attribution of mental states to other species. In *Animal Behaviour* (Vol. 109, pp. 167-176). doi.org/10.1016/j.anbehav.2015.08.011.
- [54] Adam Waytz, Carey K. Morewedge, Nicholas Epley, George

- Monteleone, Jia-Hong Gao & John T. Cacioppo, J. T. (2010). Making sense by making sentient: effectance motivation increases anthropomorphism. *Journal of personality and social psychology*, 99(3), 410. doi.org/10.1037/a0020240.
- [55] Francesco Bossi, Cesco Willemse, Jacopo Cavazza, Serena Marchesi, Vittorio Murino, Agnieszka Wykowska, 2020. The human brain reveals resting state activity patterns that are predictive of biases in attitudes toward robots, *Science robotics*, Vol.5 (46). Doi.org/ 10.1126/scirobotics.abb6652.
- [56] Nicolas Spatola, Sophie Monceau & Ludovic Ferrand, 2019. Cognitive impact of Social Robots: How anthropomorphism boosts performance. In *IEEE Robotics & Automation Magazine*, vol. 27, no. 3, pp. 73-83, Sept. 2020. doi.org/10.1109/MRA.2019.2928823.
- [57] John T. Cacioppo & Richard E. Petty, 1982. The need for cognition. *Journal of personality and social psychology*, 42(1), 116. doi.org/10.1037/0022-3514.42.1.116.
- [58] Arthur R. Cohen, Ezra Stotland & Donald M. Wolfe, 1955. An experimental investigation of need for cognition. *The Journal of Abnormal and Social Psychology*, 51(2), 291. Doi.org/10.1037/h0042761.
- [59] Athony G. Greenwald, Mahzarin R Banaji, Laurie A. Rudman, Shelly D. Farnham, Brian A. Nosek & Deborah S. Mellott, 2002. A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological review*, 109(1), 3. doi.org/10.1037/0033-295X.109.1.3
- [60] Kyveli Kompatsiari, Jairo Pérez-Osorio, Davide De Tommaso, Giorgio Metta and Agnieszka Wykowska, "Neuroscientifically-Grounded Research for Improved Human-Robot Interaction," 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, 2018, pp. 3403-3408, doi: 10.1109/IROS.2018.8594441.