International Journal of Social Robotics
https://doi.org/10.1007/s12369-022-00962-2

Body Form Modulates the Prediction of Human and Artificial Behaviour from Gaze Observation

Michele Scandola[1] · Emily S. Cross[2,3,4] · Nathan Caruana[3] · Emmanuele Tidoni[5]

Accepted: 19 December 2022

Abstract
The future of human–robot collaboration relies on people's ability to understand and predict robots' actions. The machine-like appearance of robots, as well as contextual information, may influence people's ability to anticipate the behaviour of robots. We conducted six separate experiments to investigate how spatial cues and task instructions modulate people's ability to understand what a robot is doing. Participants observed goal-directed and non-goal directed gaze shifts made by human and robot agents, as well as directional cues displayed by a triangle. We report that biasing an observer's attention, by showing just one object an agent can interact with, can improve people's ability to understand what humanoid robots will do. Crucially, this cue had no impact on people's ability to predict the upcoming behaviour of the triangle. Moreover, task instructions that focus on the visual and motor consequences of the observed gaze were found to influence mentalising abilities. We suggest that the human-like shape of an agent and its physical capabilities facilitate the prediction of an upcoming action. The reported findings expand current models of gaze perception and may have important implications for human–human and human–robot collaboration.

Keywords Gaze perception · Body perception · Action prediction · Human–robot interaction · Mentalising

1 Introduction

Inferring people's mental states from observing their gaze requires the ability to detect where they are looking and interpret the observed gaze as an expression of a desire or goal-directed behaviour [1, 2]. The engagement of higher-order social-cognitive representations during gaze-based interactions is necessary given that unlike other, more spatially-precise and unambiguous non-verbal spatial cues (e.g., hand pointing or object grasping), the communicative intent and significance of gaze shifts can be highly ambiguous [3]. A person seemingly gazing towards an apple may be signalling admiration for the apple, an intention to grasp and eat it, or may actually be looking into the space beyond the apple while they are distracted or contemplating their day. Despite the high degree of variability in the meaningfulness and communicative nature of gaze shifts, humans have a remarkable sensitivity to parse the gaze information of others so as to make predictions about a social partner's behaviour and effectively interact with them [4].

Current social models of mental state attribution from gaze perception [5] suggest a bidirectional relationship between processing visual sensory information (e.g., processing others' eye movements) and top-down cognitive influences (e.g., assuming an agent has a mind, [6, 7]). For example, the tendency to anthropomorphise non-human objects like cars predicts the activation of brain areas associated with human face processing [8]. Personal beliefs also modulate how we perceive non-human behaviour. Observing the incongruent movements of a dot or an avatar's hand when participants believe them to be controlled by a human interferes more than when participants believe the dot or virtual hand movements were programmed by a computer [9, 10]. Similarly, several

✉ Emmanuele Tidoni
e.tidoni@hull.ac.uk

1 NPSY.Lab.VR & BASIC_NPSY, Department of Human Sciences, University of Verona, Verona, Italy

2 Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, Scotland, UK

3 Department of Cognitive Science, Macquarie University, Sydney, Australia

4 MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Sydney, Australia

5 Human Technology Laboratory, Department of Psychology, University of Hull, Hull, UK

Published online: 24 January 2023

studies have now demonstrated that subjective experiences, behavioural strategies for coordination, and the neural processing of gaze differs when participants interact with virtual avatars or robots believed to be either human- or computer-controlled [7, 11, 12].

Recently, it has been suggested that processing the bodily appearance of an agent is also fundamental for social interactions, and it can inform an observer about their personality and mental states [13]. For example, Morales-Bader et al. [14] showed that people attribute more intentionality to randomly moving objects when they have a human shape compared to a triangle figure and when triangles are labelled as persons rather than mere figures. However, when we observe a directional cue that does not predict the location of a target, the form of the observed cue (e.g., eyes embedded in a human face, an apple, or a glove [15]; or an agent's tongue, [16]) does not influence the orienting of the participant's attention towards the gazed-at location.

It was originally suggested that the ability to detect an agent, that is, the identification of an entity capable of goal-directed behaviours, should not be separated from the perception of volitional actions ([1, 17]). This may imply that when detecting an agent (as defined above), its goal-directed actions should be perceived as intentional or, in other words, the observed actions should be interpreted as being guided by the intentional state of the agent. For example, when we see a friend grasp a bottle of water, we automatically perceive this behaviour as intentional, and that our friend's hand movements are willingly controlled by them. However, robots may be perceived as an agent unable to intentionally act as humans do. This implies that inferring the intentions of a human may be easier compared to reading the behaviour of a different agent (say, a robot). Moreover, such advantage should not be affected by the scenario where the two agents are acting or by changing the agent's internal state participants have to identify. In other words, the difference in the ability to read the intentions of two agents should not change as a function of contextual information (i.e., new scenarios or new task instructions) if the behaviours of the observed agents are identical across both contexts.

We recently have suggested that detecting where a humanoid robot is gazing may rely on visually processing the direction of the observed gaze (e.g., right) rather than engaging the observer's the motor system responsible for anticipating other's actions [18]. Furthermore, although participants were slower in attributing mental states to humanoid robots than to humans, we showed that the human-like appearance of non-human agents may engage processes responsible for ascribing intentions to others.

Here, we expand and replicate our previous study showing that it is easier to attribute intention to humans rather than non-human agents from the observation of non-predictive gazes in two novels laboratory and online experiments (Experiments 1 and 3 respectively). Furthermore, we investigate how what the agents gazed at (looking at a graspable object or to an empty space; Experiment 2) and task instructions (predicting what the agent is going to do vs what the agent is looking at; Experiments 4, 5, and 6) affect people's ability to interpret non-human agents with a human-like body. If the physical form of the agent is not relevant for attributing intentions to others, then changing contextual information (i.e., scenarios and task instructions) should not affect participants' ability to interpret non-human gaze behaviour. Hence, we should expect people to always be faster in attributing intentions to a human than a robot actor [18]. In contrast, if the physical form of the agent does indeed interact with contextual information, we should expect people being equally fast to attribute an intention to robots and humans in a new scenario, and an integrative account taking into consideration how an agent's bodily form may support human ability to infer others mental states through gaze observation should be considered.

2 Materials and Methods

2.1 Experiments Overview

We provide here a summary of each experiment's aims. Methodological information are in the next sections. For further details about the instructions, statistical results and data interpretation, the reader can refer to each experiment's dedicated section.

Experiment 1 was laboratory-based. The goal was to test how people infer human and robot intentions when the agent can gaze at two objects individually and randomly placed to the right or left of the agent.

Experiment 2 was conducted online and aimed at replicating Experiment 1 with only one object displayed next to the agent. We reasoned that if others' gaze is the only relevant cue to attribute intentions, we should observe similar results between Experiments 1 and 2.

Experiment 3 was conducted online and aimed at replicating the methods and results of Experiment 1 (i.e., two objects presented simultaneously, each object randomly placed to the agent's right or left).

Experiment 4 was conducted online and aimed at replicating Experiment 2. In particular, we reasoned that if participants in Experiment 2 were interpreting the agent's gaze in terms of motor goals, we should find similar results in Experiment 4 when task instructions focused on the motor consequences of the observed gaze.

Experiment 5 was conducted online and aimed at replicating Experiment 2. In particular, we reasoned that if participants in Experiment 2 were interpreting the observed gaze in terms of what the agent was perceiving, we should

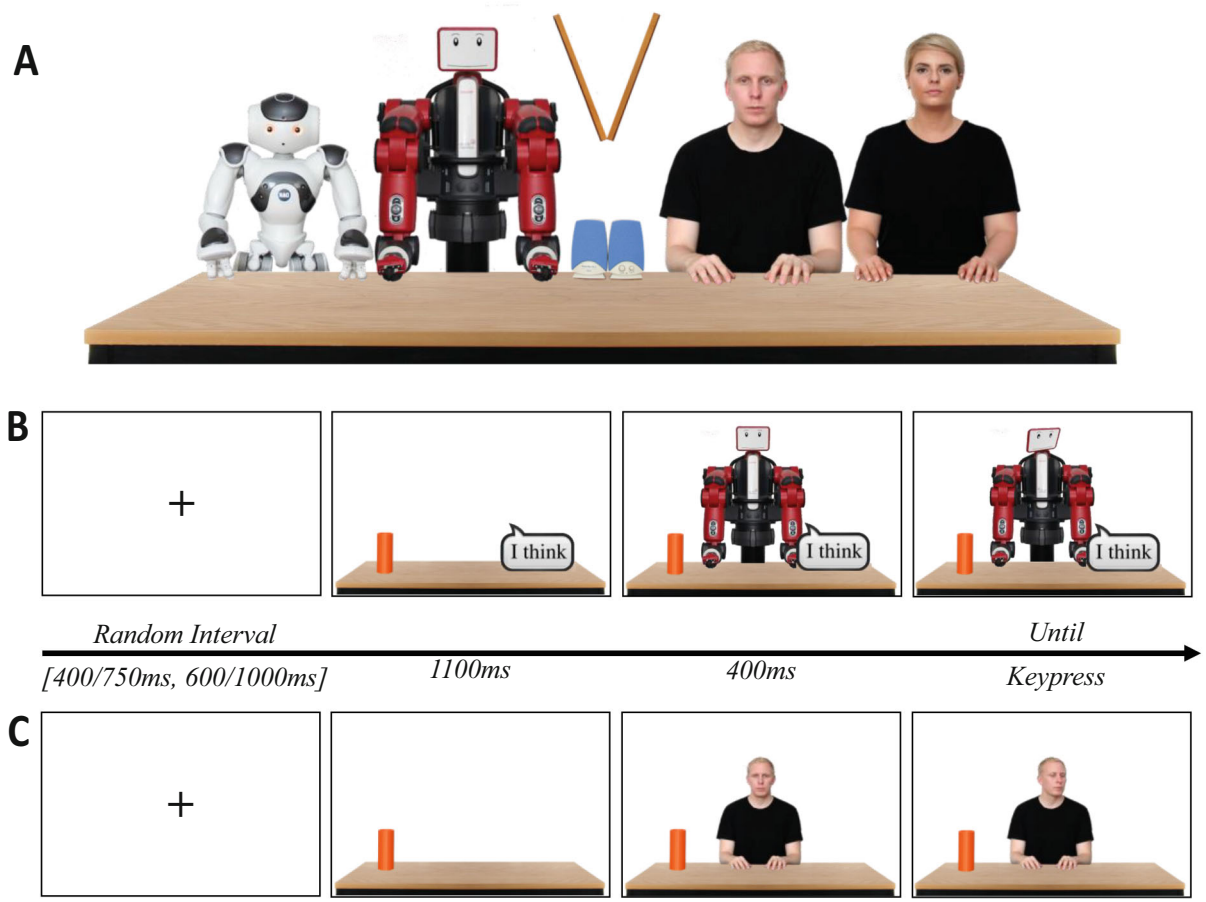

Fig. 1 The set of agents and trial timeline, *Note*. **A** Agents are displayed side by side for graphical purposes. During the experiments, agents were always centred to screen. The triangle-shaped object was presented with a set of speakers to suggest its capacity to emit sounds. **B** An example of the trial timeline: after a variable interval, the scenario is displayed and followed by the agent looking straight for 400 ms before gazing towards one direction (in the image, an example of the agent looking towards the speech bubble). **C** In Experiments 2, 4, 5, and 6, the graspable objects were the only object visible on the table. See Method section and Movie S1 for further details

find similar results in Experiment 5 when task instructions focused on the perceptual experience of the agent.

Experiment 6 was conducted online and aimed at replicating Experiment 5 findings and understanding results from Experiment 4. Specifically, we reasoned that providing participants with instructions that better define the motor and non-motor intentions guiding an agent's gaze, we should obtain similar results between Experiments 2 and 6.

2.2 General Methodology

In six separate experiments, participants observed an agent appearing behind a table and gazing towards different objects, up or down (see Apparatus and Task and Movie S1 for further details). Participants indicated either what the agent was going to do (Experiments 1–4 and 6) or what the agent was looking at (Experiment 5). Agents were two human actors (one male, one female), two humanoid robots (NAO, Softbank Robotics; Baxter, Rethink Robotics), and a triangle-shaped object with realistic visual texture (a portable lectern). Humans and humanoid robots had their trunks and upper arms visible. We edited stimuli to give the impression that their upper limbs were resting on the table (Fig. 1A). The faces of Baxter were created from an open-source database [19]. Experiments 1–3 were pre-registered, which included the fixing of sample sizes and the statistical approach taken to evaluate the data ahead of data collection for Experiments1–3 (please find the links to the online pre-registration files in the footnote[1]). Experiments 4, 5, and 6 were not pre-registered as we tried to replicate results obtained in Experiment 2 with a smaller sample size and slightly different task instructions. We adopted the same statistical approach of our previous

[1] Experiment 1: https://aspredicted.org/SUV_JQF.
Experiment 2: https://aspredicted.org/DLN_ZEK.
Experiment 3: https://aspredicted.org/MDC_LKC.

Table 1 .

Experiment	Sample	Mean age ± S.E.M., Range
1	n = 24 [female = 14, male = 10, prefer not to say = 0]	23.46 ± 1.51, [18–39]
2	n = 100 [female = 51, male = 48, prefer not to say = 1]	29.20 ± 1.08, [18–73]
3	n = 100 [female = 46, male = 53, prefer not to say = 1]	26.68 ± 0.79, [18–56]
4	n = 81 [female = 25, male = 54, prefer not to say = 2]	25.06 ± 0.83, [18–65]
5	n = 80 [female = 24, male = 55, prefer not to say = 1]	24.39 ± 0.55, [18–44]
6	n = 81 [female = 28, male = 53, prefer not to say = 0]	25.44 ± 0.75, [18–53]

Descriptive measures of the sample for each experiment

work [18] to facilitate comparison across all experiments and studies.

2.3 Participants

We recruited adult English-speaking participants from the University of Hull (Experiment 1) and via the online research participation platform Prolific Academic [20] (Experiments 2–6) in exchange for course credits and monetary compensation, respectively. The task, procedure, and methodology were reviewed and approved by the institutional review boards of the University of Hull (protocol number: FHS150) and carried out following the standards set by the Declaration of Helsinki (2013). All participants were naïve to the task and purpose of the experiment. Each participant completed a single experiment, and informed consent was obtained before starting the task. Participants were free to withdraw from the experiment any time without having to give any reason (for online experiments participants were informed they could withdraw by pressing the 'escape' key).

A total of 465 participants completed different experiments (see Table 1). Sample sizes were selected (G*Power, [21]) so that we had sufficient power to detect medium to large effect sizes for Experiments 1 (dz = 0.6, Alpha = 0.05, Beta = 0.80), small to large effects for Experiments 2 and 3 (dz = 0.3, Alpha = 0.05, Beta = 0.80). We slightly reduced sample sizes for Experiments 4, 5, and 6 due to limited resources. However, our sample sizes were nonetheless sufficient to detect small to large effect sizes (dz = 0.35, Alpha = 0.05, Beta = 0.80).

2.4 Procedure

Participants were invited to read the information sheet and communicate any questions to the experimenter if needed. After providing informed consent, participants were explained the experimental task in the laboratory Experiment 1 and read the experimental instructions in the other online experiments. Three agents were presented and described as humans, robots, and a triangle capable of performing three actions (i.e., a total of 9 experimental conditions), and the trial timeline was explained (see Movie S1 for instructions and trial timeline). After that, participants performed a quick practice session of 18 trials to ensure participants could match the observed action with the corresponding key. Since online data collection was performed without the direct supervision of the experimenter, the online practice session provided accuracy feedback to participants after their response for the first 12 practice trials (see Movie S1). After the practice session, participants started four experimental blocks. We decided to split the experimental session into four blocks to allow participants to take some breaks. Each block comprised 54 trials for Experiment 1 (a total of 216 trials; 24 trials per condition) and 45 trials for Experiments 2–6 (a total of 180 trials; 20 trials per condition). Agents' presentation was pseudorandomised across the four experimental blocks (i.e., each trial may have displayed a different agent). After the main task, participants rated their exposure to media robotic content ("How often do you watch movies, TV series, or play videogames where robots are involved?") using a nominal scale (1 = Never, 2 = Once every Year, 3 = Once every 6 months, 4 = Once every 3 months, 5 = Once every month, 6 = More than once every month). After the experiment, participants were debriefed as to the purpose of the experiment.

In Experiment 1, participants also completed a computerised series of questionnaires related to attitudes and perception of robots. In particular, they reported their level of agreement along a horizontal bar to the items of the Negative Attitudes towards Robots Scale (NARS; [22]) and the Robot Anxiety Scale (RAS, [22]). Moreover, participants answered to some items of the anthropomorphism, animacy, and intelligence subscales of the Godspeed questionnaire [23] by indicating their position on a scale between two bipolar words (e.g., human-like, machine-like). Finally, we assessed participants' opinions about how close they perceive themselves and other humans to robots using a modified and computerised version of the Inclusion of the Other in the Self [24]. Questionnaires (see Supplementary Table 1) were collected to have insights about the perception of robots within the sample and were not analysed further.

2.5 Apparatus and Task

We used 3D printed geometrical shapes as graspable objects (i.e., a cube, a cylinder, a sphere, and a rectangle). In the laboratory Experiment 1, participants were allowed to manipulate the graspable objects as long as they liked before the task (naturally, this was not possible for online experiments). The task for Experiment 1 was designed using Matlab 2018R and Psychotoolbox 3 [25]. Stimuli were presented on a 21.5-inch screen (resolution: 1920 × 1080 pixels; refresh frequency, 60 Hz) at a distance of 70 cm from the participant. For online experiments, the scripts were designed using Psychopy 3 [26] and hosted on Pavlovia (Pavlovia.org) to run within the participants' browser.

In Experiments 1 and 3, each agent could gaze towards a graspable object, a text bubble, or up and down (Fig. 1B). We displayed only one of the four graspable objects in each trial and the text bubble could contain one of ten short self-descriptive sentences (i.e., I think, I plan, I desire, I judge, I worry, I believe, I imagine, I relax, I feel, I like). These short sentences were selected to facilitate both physical and mental states attribution to the observed agent [27]. Graspable objects and sentences were not associated with a specific agent and were randomly presented each trial. In Experiments 2, 4, 5, and 6, the text bubble was removed, and the agent appeared to gaze at the empty table (Fig. 1C). The location of the graspable object and text bubble (or the empty table for Experiments 2, 4, 5, and 6) was randomly generated for each trial. Participants were asked to place their right index, middle, and ring fingers over three keys ('n', 'j', 'i'). These kyes were chosen because they allow a normal and relaxed position of the right arm and fingers. Keys were randomly assigned to one action across participants. In all Experiments, the up and down gaze movements were associated with the same (randomised across participants) key.

The trial timeline was identical in all Experiments, with few differences between laboratory and online experiments. Participants observed the scenario for 1500 ms in Experiments 1 and 1100 ms in all other online Experiments. Then, one of the three agents could appear, and 400 ms later they turned their heads and gazed right or left (towards the graspable object and text bubble for Experiments 1–3; towards the graspable object and the empty table for Experiments 2, 4, 5, and 6) or either up or down (for all Experiments). This rapid succession of the agents' images created the impression of an apparent motion [28, 29]. Finally, the agent and the environment remained visible until response. In Experiment 1, the intertrial interval randomly ranged between 750 and 1000 ms. In the online experiments, the intertrial interval randomly ranged between 400 and 600 ms.

2.6 Measures, Data Processing, and Statistical Approach

We collected task Accuracy and Response Time (RT; expressed in seconds) as performance measures, and we specified how data would be processed in the online pre-registration files. Specifically, for all experiments, we excluded trials < 0.150 s and > 1.500 s. Then, we excluded trials whose RTs fell above or below 2.5SDs of the overall mean within each block for each participant. At this stage of data processing, we excluded participants whose overall accuracy was below 65%. Finally, we excluded participants whose performance (in RTs or Accuracy) fell above or below 2.5SDs of the overall mean across conditions of the remaining participants.

Statistics were performed using R 3.5.1 [30] run on the University of Hull High-Performance facility VIPER (http://hpc.wordpress.hull.ac.uk/home/). We used the lme4 package (v1.1.27.1; [31]) to perform MLM with fixed effects and complex random intercepts (CRIs) as random effects.[2] Model reduction started from the full-CRIs MLM[3]. For MLMs on RT of correct answers, we also report the partial eta-squared as a measure of effect size (effectsize v0.4.5; [32]). For all MLMs, we computed the conditional R2 (for lme4::lmer performance v0.7.3, [33]; for lme4::glmer MuMIn v1.43.17, [34]). Throughout the paper, we report the p-values computed on the estimates of the simplified MLM. For each multiple comparison, we report the individual Bonferroni corrected p-value computed from the final MLM using emmeans (v1.6.2-1; [35]). Furthermore, we performed confirmatory ANOVAs on Accuracy and mean-aggregated RT data to support the main analyses. For each confirmatory analysis, we ran multiple comparisons[3] and reported the absolute value of the Cohen's d (|d|) and the Bayes Factor (BF10; default Cauchy prior of 0.707; JASP Team, 2021, Version 0.14) to further facilitate the reader in assessing the strength of the evidence. Classically, BF10 is interpreted as showing very strong evidence towards the alternative hypothesis when greater than 150, strong evidence when equal or greater than 20, positive evidence when equal or greater than 3, and with weak or negligible evidence when between 1 and 3 [36]. The inverse of these values (1/150, 1/20, 1/3) can be interpreted as BF10 showing very strong, strong, or positive evidence towards the null hypothesis.

We considered as non-conclusive discordant findings obtained from the MLM and the analyses on mean-aggregated data. If not stated otherwise, the ANOVAs and

[2] Scandola, M., **Tidoni, E.**, (*Preprint*). The development of a standard procedure for the optimal reliability-feasibility trade-off in Multilevel Linear Models analyses in Psychology and Neuroscience. https://psyarxiv.com/kfhgv.

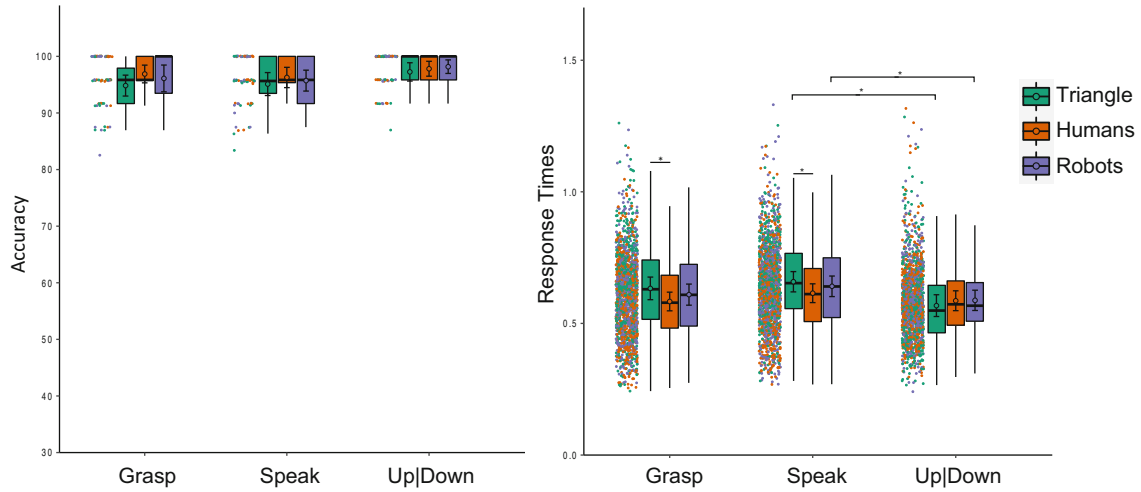

Fig. 2 Results of experiment 1, *Note*. Participants' performance (Accuracy percentage on the left, Response Times expressed in seconds on the right) in Experiment 1 for each experimental condition. To provide a comprehensive overview of collected data, raw data from each experimental condition are visualised as raincloud plots, median bar plots (with lower and upper hinges corresponding to the 25th and 75th percentile and whiskers extending no further than 1.5 * "Inter Quartile Range" from the hinge), and probability density. The circles inside each median bar plot indicate the average of the by-subject mean-aggregated data for that condition. Error bars represent 95% confidence intervals of the mean based on subject-aggregated data. Data visualisation has been possible by adapting the open-source R code "RainCloudPlots" [37]. Asterisks denote the significant differences ($p < 0.05$) for both the MLM and the ANOVA on mean-aggregated data as reported in the main text. Section sign symbol (§) denotes a tendency (i.e., a significant p-value for the MLM but a p-value comprised between 0.05 and 0.10 in the ANOVA on mean-aggregated data)

multiple comparisons confirmed the results obtained from the MLM model.

3 Experiment 1

Participants (n = 24) were instructed that each agent could look towards the graspable object to grasp it (gaze to grasp: "is going to grasp the object"), towards the text bubble to speak (gaze to speak: "is going to speak"), up or down to do nothing (non-goal directed action as control: "is looking up or down"). Participants were asked to indicate what the agent was going to do (i.e., the agent is going to grasp the object, going to speak, or is looking up or down). If inferring what an agent is going to do is easier when we observe human agents, we should expect faster RT when humans but not the other agents look towards an object. No differences across agents are expected when they perform a non-goal directed gaze.

We removed trials with RTs deemed too fast or too slow (0.39%). Then, trials with RTs falling above or below 2.5SD of the overall mean within each block of each participant were removed (1.78%). No participants' performance was < 65%. Finally, one participant had a performance above 2.5SD of the overall mean across conditions of the remaining participants and was excluded from the final sample (n = 23).

3.1 Results

We analysed performance measures (see Fig. 2) with Agent (human, robot, triangle) and Action (to grasp, to speak, look up|down) as within-subject fixed effects of a multilevel linear model (MLM; see Table S2 in Supplementary Materials for details on the fixed and random effects structure of all the MLMs). In the case of a two-way Agent by Action interaction, we performed eighteen multiple paired comparisons of interest. Specifically, we compared the three Action levels within each Agent (e.g., graspable object—human agent vs text bubble—human agent; 9 comparisons), and each Action level across the three agents (e.g., text bubble—human agent vs text bubble—robot agent; 9 comparisons). We also performed a confirmatory ANOVA on Accuracy and mean-aggregated RT data to support the main analyses. Non-conclusive findings (i.e., less robust, more fragile) are highlighted whenever the MLM and the confirmatory analyses yield discordant results (see Method section for further details).

For Accuracy, we observed a main effect of Action, $\chi^2(2) = 7.836$, $p = 0.020$. This was not confirmed by the confirmatory ANOVA ($F = 3.430$, $p = 0.054$, $\eta p2 = 0.135$). We observed no main effect of Agent, $\chi^2(2) = 3.671$, $p = 0.160$, and no significant Gaze by Agent interaction, $\chi^2(4) = 0.844$, $p = 0.932$.

For RT, we removed incorrect answers (3.53%) from the final dataset. We observed a main effect of Agent, $F(2,44) = 11.668$, $p < 0.001$, $\eta p2 = 0.349$, a main effect of Action, $F(2,44) = 6.787$, $p = 0.003$, $\eta p2 = 0.236$, and a significant Action by Agent interaction, $F(4, 88) = 8.703$, $p < 0.001$, $\eta p2 = 0.286$. Interested readers can find supplementary descriptive statistics and statistical comparisons of the main effects for all experiments via the Open Science Framework (https://osf.io/bd6h3/). The latter suggested that participants were faster in detecting the triangle looking up|down (0.568 ± 0.020 s) compared to attributing the intention to speak (0.658 ± 0.019 s; $p < 0.001$, $|d| = 1.207$, $BF_{10} = 2.659e + 03$) and to grasp (0.633 ± 0.021 s; $p = 0.006$). However, the latter did not survive Bonferroni correction using pairwise t-tests on aggregated data ($p = 0.167$, $|d| = 0.595$, $BF_{10} = 5.222$). Participants tended to be faster in detecting the robot looking up|down (0.588 ± 0.019 s) compared to attributing the intention to speak (0.641 ± 0.019 s; MLM $p = 0.060$, Confirmatory ANOVA $p = 0.028$, $|d| = 0.752$, $BF_{10} = 24.031$).

Furthermore, participants were faster in attributing to humans the intention to grasp (0.583 ± 0.017 s) compared to the triangle ($p < 0.001$, $|d| = 1.232$, $BF_{10} = 3.456e + 03$). A similar tendency was observed comparing the attribution of the intention to grasp to humans and robots (0.609 ± 0.019 s; MLM $p = 0.079$, Confirmatory ANOVA $p = 0.040$, $|d| = 0.722$, $BF_{10} = 17.745$). Finally, RT was faster in attributing the intention to speak to humans (0.615 ± 0.017 s) compared to the triangle (0.658 ± 0.019 s; $p < 0.001$, $|d| = 0.734$, $BF_{10} = 19.992$). No other Bonferroni corrected p-values were lower than 0.05 for both multiple comparisons computed on the estimates of the simplified MLM and confirmatory multiple comparisons using pairwise t-tests ($p > 0.132$, $|d| < 0.529$, $BF_{10} < 2.915$; see Methods for data analysis approach).

3.2 Discussion Experiment 1

We observed that people are faster to infer the next action of an agent when observing humans compared to non-humanlike agents. Participants were faster to attribute the intention to grasp and speak to humans compared to the triangle. Moreover, participants tended to be faster at detecting the intention to grasp among humans compared to robots. We observed no difference when the agents did not look towards an object (up|down condition).

Notably, we observed that attributing the intention to speak took longer for the triangle and robots compared to the simple detection of up|down movements. Such differences were absent for humans. This pattern of findings may suggest that participants implicitly or explicitly considered the nature of the observed agent, and reflected upon their mental states rather than using a visual-only strategy to solve the task. Moreover, while RTs in attributing a motor and a social intention to humans were faster compared to the triangle, RTs in attributing motor and social intentions to robots did not differ from humans and the triangle. This may suggest that robots were perceived as neither human nor as non-human agents. However, we observed a bigger difference between humans and robots for attributing the intention to grasp compared to the intention to speak, as inferred by the effect sizes, and by the fact that the BF10 for the speak intention was non-conclusive (grasp: $|d| = 0.722$, BF10 = 17.745; speak: $|d| = 0.529$, BF10 = 2.915). This result may indicate that participants have processed motor and communicative intentions differently or that the prediction of hand motor actions favoured faster manual responses for the grasping rather than the speaking intention. However, such potential contributions of the motor system in predicting a manual action compared to a non-manual one facilitated only the participants' interpretation of the human gaze, and did not extend to robotic gaze (i.e., robots did not differ from the triangle).

Overall, the findings from Experiment 1 suggest that when participants are asked to reflect upon others' mental content, the interpretation of the observed action is faster for humans compared to non-humanlike agents. Interestingly, RTs performance suggests that robots may have been perceived as hybrid agents, part human, part non-human.

4 Experiment 2

The first experiment showed that attributing intentions to humans required less effort than attributing the same intention to robots and inanimate objects. In Experiment 2, we tested if a similar pattern of results is observed after removing the text bubble from the table and biasing participants' attention towards the graspable object. If attributing intentions varies exclusively based on the cues provided by the new scenario, then we should expect a change in the general performance from Experiment 1 (i.e., faster RT when agents gaze towards the graspable object than the empty table) but no changes depending on the observed agent (i.e., triangle and robots should still not differ).

Participants (n = 100) were instructed that each agent could look towards the graspable object to grasp it ("is looking at the object to grasp"), towards the empty table ("is looking at the empty table"), up or down ("is looking up or down"). Note that we slightly changed how we phrased the instructions for the gaze towards the graspable object from Experiment 1 (Experiment 1 instruction: "is going to grasp the object") to match the new non-motor condition when agents looked at the empty table. Participants were asked to indicate what the agent was going to do (i.e., the agent is looking at the object to grasp, at the empty table, up or down). We removed trials with RTs deemed too fast or too slow (8.53%). Then, trials with RTs falling above or below 2.5SD of the overall mean within each block of each

participant were removed (1.97%). Three participants' performance was < 65%. Seven participants with a performance above 2.5SD of the overall mean across conditions of the remaining participants were excluded. Despite this data management approach, one participant had 0% accuracy when the human looked to the graspable object, suggesting a misunderstanding of the task. Thus, we removed that participant (final sample n = 89).

4.1 Results

We analysed performance measures (see Fig. 3) with Agent (human, robot, triangle) and Action (to grasp, to look at the table, look up|down) as within-subject fixed effects of the MLM. In the case of a two-way Agent by Action interaction, we performed eighteen multiple paired comparisons of interest as indicated in Experiment 1.

For accuracy, we observed a main effect of Action, $\chi^2(2) = 20.798$, $p < 0.001$. Participants were less accurate in attributing the intention to look at the empty table (95.23 ± 0.56%) compared to detecting a non-goal directed action (97.48 ± 0.42%; $p < 0.001$, |d|= 0.438, $BF_{10} = 230.804$). The difference between the non-goal-directed action and the intention to grasp (96.18 ± 0.44%; $p = 0.004$) was not confirmed using Bonferroni corrected pairwise t-tests on aggregated data ($p = 0.102$, |d|= 0.228, $BF_{10} = 1.050$). We observed no main effect of Agent, $\chi^2(2) = 1.057$, $p = 0.589$, and no significant Gaze by Agent interaction, $\chi^2(4) = 3.252$, $p = 0.517$.

For RT, we removed incorrect answers (3.67%) from the final dataset. We observed a main effect of Agent, $F(2524) = 32.071$, $p < 0.001$, $\eta p2 = 0.113$, a main effect of Action, $F(2176) = 56.187$, $p < 0.001$, $\eta p2 = 0.389$, and a significant Action by Agent interaction, $F(4523) = 4.962$, $p < 0.001$, $\eta p2 = 0.038$. The latter suggested that participants were faster in detecting the triangle looking up|down (0.709 ± 0.014 s) compared to attributing the intention to grasp (0.752 ± 0.015 s; $p < 0.001$, |d|= 0.380, $BF_{10} = 39.483$) and to look at the table (0.796 ± 0.013 s; $p < 0.001$, |d|= 0.976, $BF_{10} = 4.624e + 11$). Participants were also slower in detecting the triangle looking at the table compared to attributing the intention to grasp ($p < 0.001$, |d|= 0.475, $BF_{10} = 782.991$). Participants were slower in detecting the human looking at the table (0.766 ± 0.014 s) compared to the human looking up|down (0.692 ± 0.012 s; $p < 0.001$, |d|= 0.866, $BF_{10} = 3.906e + 09$) and gazing at the graspable object (0.694 ± 0.015 s; $p < 0.001$, |d|= 0.822, $BF_{10} = 6.064e + 08$). Participants were also slower in detecting the robot looking at the table (0.784 ± 0.014 s) compared to the robot looking up|down (0.705 ± 0.013 s; $p < 0.001$, |d|= 0.885, $BF_{10} = 9.015e + 09$) and gazing at the graspable object (0.714 ± 0.014 s; $p < 0.001$, |d|= 0.797, $BF_{10} = 2.031e + 08$).

Furthermore, participants were slower to attribute the intention to grasp to the triangle compared to humans ($p < 0.001$, |d|= 0.725, $BF_{10} = 1.022e + 07$) and robots ($p < 0.001$, |d|= 0.561, $BF_{10} = 1.600e + 04$). Finally, participants were faster to attribute the intention to look at the table to humans compared to the triangle ($p = 0.003$, |d|= 0.344, $BF_{10} = 14.744$). No other Bonferroni-corrected p-values were lower than 0.05 for both multiple comparisons computed on the estimates of the simplified MLM and multiple comparisons using pairwise t-tests on aggregated data ($p > 0.150$, |d|< 0.312, $BF_{10} < 6.467$).

4.2 Discussion Experiment 2

We observed that participants were slower to predict goal-directed actions compared to non-goal directed actions when observing the triangle. Contrary, RTs for attributing the intention to grasp did not differ from RTs of the control condition for human and robotic agents. Moreover, we observed slower RTs for the triangle when pointing towards the graspable objects. These results cannot be explained by a participants' attentional bias towards the side of the screen where graspable objects were displayed [38]. Contrary, these findings may suggest that the motor acts evoked by the simple observation of the graspable objects [39, 40] had to match the observed agent's motor capabilities. This leads to the idea that human and robotic physical appearance may have facilitated the perception of those agents as capable of performing grasping actions or increased the perception of human-like agency traits (e.g., the ability to think, plan, see). In this sense, both humans and robots differed from the triangle in attributing the intention to grasp, but only the human differed from the triangle in attributing the intention to look at the table. This may suggest that perceptual information derived from analysing the shape and visual texture of the agent, like its physical and mental capabilities, may have played a major role. In other words, both robots and humans can move, and they may be expected to act on an object. However, observing a gaze towards the table might have implied the ability of the agent to see and represent the external world. That is, asking participants what an agent is looking at may have facilitated the mental state attribution of seeing [6, 41, 42] to humans more than robots.

However, based on the findings from Experiment 1, some of the results from Experiment 2 were unexpected. For example, we expected a difference between humans and robots in attributing the intention to grasp and no difference between the robot and the triangle. For this reason, we wanted to ensure that the findings of Experiment 1 were not due to differences between laboratory and online settings or due to the different sample size between Experiment 1 and Experiment 2. Hence, our next step was to replicate Experiment 1 online.

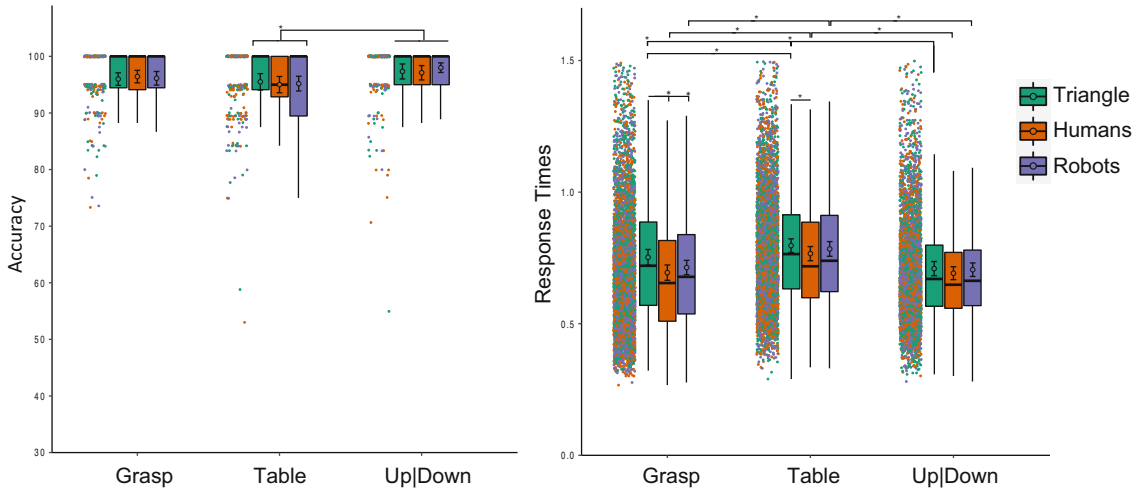

Fig. 3 Results of experiment 2, *Note.* Participants' performance in Experiment 2. The labels "Grasp" and "Look" indicate the conditions where participants attributed to the observed agent the intention to grasp and look at the empty table, respectively. See Fig. 2 for a detailed explanation of our data visualisation approach

5 Experiment 3

Experiment 2 showed that attributing motor intentions to humans and robots required less effort than attributing the intention to grasp to the triangle. In Experiment 3, we showed both the graspable objects and the text bubble on the table to replicate Experiment 1 online.

Participants (n = 100) were instructed that each agent could look towards the graspable object to grasp it ("is going to grasp the object"), towards the speech bubble ("is going to speak"), up or down ("is looking up or down"). Participants were asked to indicate what the agent was going to do (i.e., the agent is going to grasp the object, going to speak, or is looking up or down).

We removed trials with RTs deemed too fast or too slow (5.92%). Then, trials with RTs falling above or below 2.5SD of the overall mean within each block of each participant were removed (2.03%). Four participants' performance was < 65%. Five participants with a performance above 2.5SD of the overall mean across conditions of the remaining participants were excluded. Despite this data management approach, one participant had 0% accuracy when the triangle looked up or down, suggesting a misunderstanding of the task. Thus, we removed that participant (final sample n = 90).

5.1 Results

We analysed performance measures (see Fig. 4) with Agent (human, robot, triangle) and Action (to grasp, to speak, up|down) as within-subject fixed effects of an MLM. In the case of a two-way Agent by Action interaction, we performed eighteen multiple paired comparisons of interest as indicated in Experiment 1.

For accuracy, we observed a main effect of Action, $\chi^2(2) = 25.918$, $p < 0.001$. Participants were more accurate in detecting a non-goal directed action (97.82 ± 0.42%) than attributing the intention to speak (95.53 ± 0.61%; $p < 0.001$, |d|= 0.390, $BF_{10} = 56.367$) and grasp (96.09 ± 0.46%; $p < 0.001$, |d|= 0.323, $BF_{10} = 8.876$). We observed no main effect of Agent, $\chi^2(2) = 0.092$, $p = 0.955$, and no significant Action by Agent interaction, $\chi^2(4) = 0.332$, $p = 0.988$.

For RT, we removed incorrect answers (3.44%) from the final dataset. We observed a main effect of Agent, F(2,531) = 43.794, $p < 0.001$, ηp2 = 0.146, a main effect of Action, F(2,178) = 77.608, $p < 0.001$, ηp2 = 0.472, and a significant Action by Agent interaction, F(4,530) = 6.620, $p < 0.001$, ηp2 = 0.049. The latter suggested that participants were faster to recognise the triangle looking up|down (0.658 ± 0.011 s) compared to attributing the intention to grasp (0.748 ± 0.012 s; $p < 0.001$, |d|= 0.988, $BF_{10} = 1.098e + 12$) and speak (0.751 ± 0.013 s; $p < 0.001$, |d|= 0.858, $BF_{10} = 3.718e + 09$). Participants were also faster to recognise the humans looking up|down (0.650 ± 0.010 s) compared to attributing the intention to grasp (0.701 ± 0.011 s; $p < 0.001$, |d|= 0.648, $BF_{10} = 5.370e + 05$) and speak (0.708 ± 0.013 s; $p < 0.001$, |d|= 0.724, $BF_{10} = 1.211e + 07$). Participants were faster to recognise robots looking up|down (0.659 ± 0.010 s) compared to attributing the intention to grasp (0.742 ± 0.013 s; $p < 0.001$, |d|= 0.995, $BF_{10} = 1.466e + 12$) and speak (0.739 ± 0.013 s; $p < 0.001$, |d|= 0.943, $BF_{10} = 1.514e + 11$).

Furthermore, participants were faster in attributing the intention to grasp to the human compared to the triangle

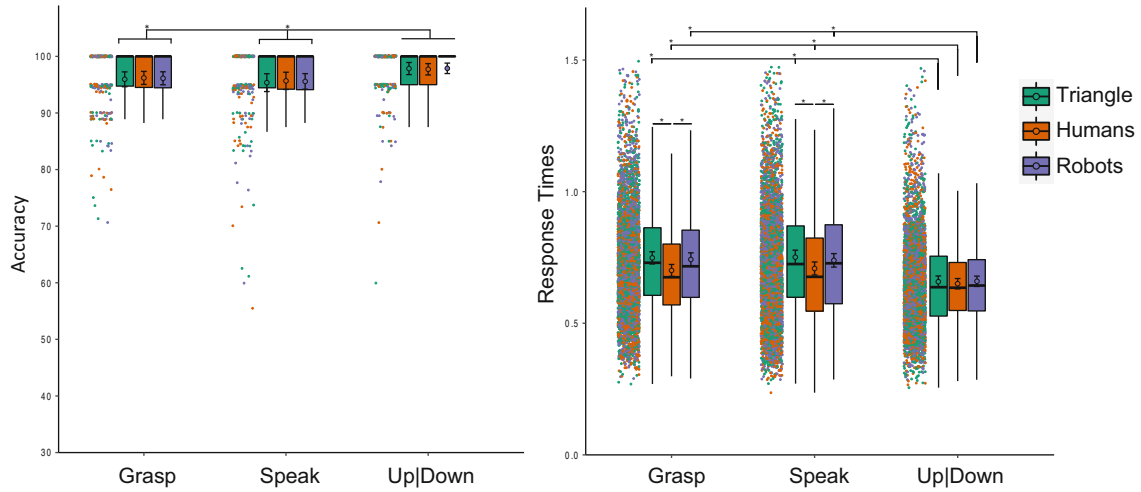

Fig. 4 Results of experiment 3, *Note.* Participants' performance in Experiment 3. The labels "Grasp" and "Speak" indicate the conditions where participants attributed to the observed agent the intention to grasp and speak, respectively. See Fig. 2 for a detailed explanation of our data visualisation approach

($p < 0.001$, |d|= 0.898, $BF_{10} = 2.065e + 10$) and robots ($p < 0.001$, |d|= 0.728, $BF_{10} = 1.441e + 07$). Participants were also faster in attributing the intention to speak to the human compared to the triangle ($p < 0.001$, |d|= 0.576, $BF_{10} = 3.289e + 04$) and robots ($p < 0.001$, |d|= 0.481, $BF_{10} = 1.038e + 03$).

No other Bonferroni corrected p-values were lower than 0.05 for both multiple comparisons computed on the estimates of the simplified MLM and multiple comparisons using pairwise t-tests on aggregated data ($p > 0.900$, |d|< 0.182, $BF_{10} < 0.486$).

5.2 Discussion Experiment 3

We replicated the findings from Experiment 1 with faster RTs when participants attribute intentions to humans rather than the triangle and robots. Moreover, these results expand previous reports where online participants were asked to infer the agents' intentions, and the graspable objects and text bubble were in a fixed position (Experiment 4 in [18]). The findings from Experiment 3 also enable us to exclude the possibility that the results in Experiment 2 were driven by a difference between laboratory and online samples.

However, Experiments 1–3 had slightly different instructions compared to Experiment 2, and may have favoured a motor interpretation of the observed gaze: instructions asked participants to indicate whether the agent was going to grasp or speak. On the contrary, in Experiment 2, instructions may have favoured also the use of a visual strategy: participants indicated whether the agent looked at the object to grasp or looked at the empty table. Hence, we manipulated the instructions in three additional experiments to investigate further the extent to which participants adopt a visual-only or motor-only strategy to solve the task.

6 Experiment 4

The previous experiments showed that attributing motor intentions to humans requires less effort than attributing intentions to non-human agents. However, when only the graspable objects were displayed (Experiment 2), no difference was observed between humans and robots. Here, we tried to replicate the findings of Experiment 2 by showing only the graspable objects and by focusing on motor instructions to solve the task. If in Experiment 2 participants adopt a motor strategy, we should expect similar results.

Participants (n = 81) were instructed that each agent could look towards the graspable object to grasp it ("is going to grasp the object"), towards the empty table ("is going to do nothing"), up or down ("is looking up or down"). Participants were asked to indicate what the agent was going to do (i.e., the agent is going to grasp the object, do nothing, or is looking up or down).

We removed trials with RTs deemed too fast or too slow (7.24%). Then, trials with RTs falling above or below 2.5SD of the overall mean within each block of each participant were removed (2.03%). Three participants with a performance < 65% were removed. Five participants with a performance above 2.5SD of the overall mean across conditions of the remaining participants were excluded (final sample n = 73).

6.1 Results

We analysed performance measures (see Fig. 5) with Agent (human, robot, triangle) and Action (to grasp, to do nothing,

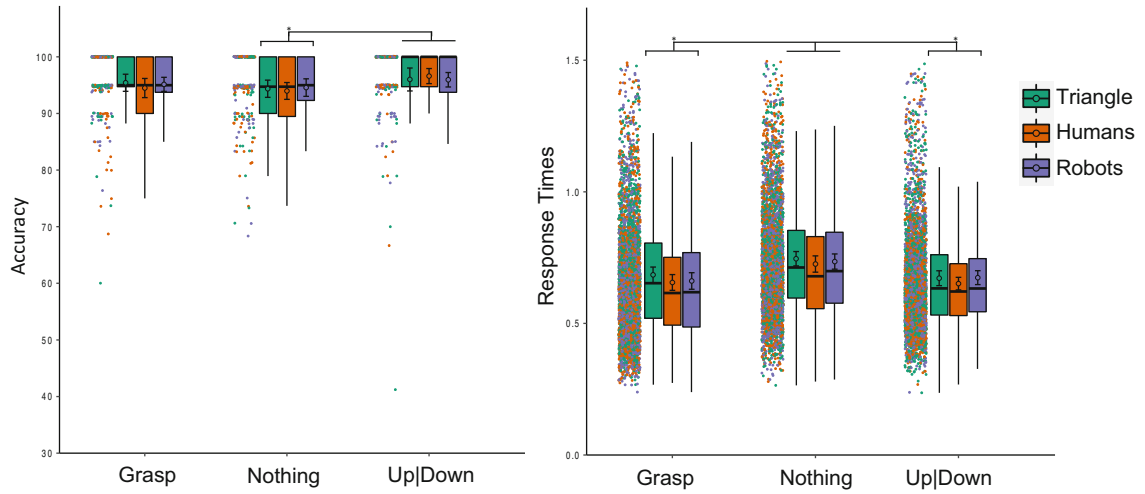

Fig. 5 Results of experiment 4, *Note.* Participants' performance in Experiment 4. No interaction between the agents and the observed action was observed. For consistency across the figures of all experiments, we display all experimental conditions. For RT, we display the main effect of Action. We invite the reader to refer to the main text for the main effect of Agent in the RTs. The labels "Grasp" and "Nothing" indicate the conditions where participants attributed to the observed agent the intention to grasp and do nothing, respectively. See Fig. 2 for a detailed explanation of our data visualisation approach

up|down) as within-subject fixed effects of an MLM. In the case of a two-way Agent by Action interaction, we performed eighteen multiple paired comparisons of interest as indicated in Experiment 1.

For accuracy, we observed a main effect of Action, $\chi^2(2) = 14.186$, $p < 0.001$. Participants were more accurate in detecting the agent looking up|down (96.19 ± 0.56%) than attributing the intention to do nothing (94.31 ± 0.54%; $p = 0.001$, |d|= 0.350, $BF_{10} = 7.510$) and to grasp (95.02 ± 0.54%; $p = 0.036$). The latter was not confirmed using Bonferroni corrected pairwise t-tests on aggregated data ($p = 0.276$, |d|= 0.200, $BF_{10} = 0.511$). We observed no main effect of Agent, $\chi^2(2) = 0.167$, p = 0.920, and no significant Action by Agent interaction, $\chi^2(4) = 1.999$, $p = 0.736$.

For RT, we removed incorrect answers (4.73%) from the final dataset. We observed a main effect of Agent, $F(2430) = 13.219$, $p < 0.001$, $\eta p2 = 0.061$, a main effect of Action, $F(2144) = 55.201$, $p < 0.001$, $\eta p2 = 0.435$, and a non-significant Action by Agent interaction, $F(4430) = 1.701$, $p = 0.149$, $\eta p2 = 0.016$.

The main effect of Action revealed that participants were slower in attributing the intention to do nothing (0.735 ± 0.014 s) compared to the intention to grasp (0.667 ± 0.014 s; $p < 0.001$, |d|= 1.204, $BF_{10} = 7.500e + 12$) or look up|down (0.666 ± 0.012 s; $p = 0.025$, |d|= 1.121, $BF_{10} = 4.010e + 11$). We observed no difference between attributing the intention to grasp and to look up|down ($p = 1.000$, |d|= 0.022, $BF_{10} = 0.131$). This result may confirm that displaying only one object may have captured participants attention and facilitated the processing of the gaze directed towards a graspable object.

The main effect of Agent revealed that participants were faster in discriminating the different gaze behaviour for humans (0.677 ± 0.013 s) compared to robots (0.690 ± 0.013 s; $p = 0.019$, |d|= 0.315, $BF_{10} = 3.611$) and the triangle (0.701 ± 0.013 s; $p < 0.001$, |d|= 0.528, $BF_{10} = 754.594$). We observed a tendency for RT to be faster for robots compared to the triangle ($p = 0.051$, |d|= 0.262, $BF_{10} = 1.325$).

A closer inspection to the experimental conditions suggests that these differences were mainly driven by faster RTs when the humans and robots gazed towards the graspable object. Indeed, exploratory Bonferroni-corrected comparisons suggested responses to the triangle gaze tended to be slower (0.685 ± 0.015 s) than responses to humans (0.656 ± 0.015 s; MLM $p = 0.002$, Confirmatory ANOVA $p = 0.059$, |d|= 0.356, $BF_{10} = 8.631$) and robots (0.661 ± 0.016 s; $p = 0.025$, Confirmatory ANOVA $p = 0.127$, |d|= 0.325, $BF_{10} = 4.387$; see Supplementary Table S3).

6.2 Discussion Experiment 4

Although the numerical trend of the findings was similar to results observed in Experiment 2, we did not fully replicate Experiment 2. Using motor instructions (doing vs not-doing) may have changed the way participants interpreted the agents' behaviour. For this reason, the next experiment we performed had instructions that focused on the non-motor aspects of gazing.

7 Experiment 5

Here, we aimed to replicate the findings of Experiment 2 by showing only the graspable objects and by focusing on visual instructions to solve the task. If in Experiment 2 participants adopted a visual strategy, we should expect similar results.

Participants (n = 80) were instructed that each agent could look towards the object ("is looking at the object"), towards the table ("is looking at the table"), look up or look down ("is looking up or down"). Participants were asked to indicate what the agent was looking at (i.e., the agent is looking at the object, at the table, or is looking up or down). Graspable objects were labelled as 'objects', and terms like 'grasping' or 'graspable objects' were not mentioned.

We removed trials with RTs deemed too fast or too slow (9.08%). Then, trials with RTs falling above or below 2.5SD of the overall mean within each block of each participant were removed (2.18%). Two participants' performance was < 65%. Seven participants with a performance above 2.5SD of the overall mean across conditions of the remaining participants were also excluded (final sample n = 71).

7.1 Results

We analysed performance measures (see Fig. 6) with Agent (human, robot, triangle) and Action (to look at the object, to look at the table, look up|down) as within-subject fixed effects of an MLM. In the case of a two-way Agent by Action interaction, we performed eighteen multiple paired comparisons of interest as indicated in Experiment 1.

For accuracy, we observed a main effect of Action, $\chi^2(2) = 12.487$, $p = 0.002$. This was not confirmed on the confirmatory ANOVA on aggregated data ($F = 3.277$, $p = 0.056$, $\eta p2 = 0.045$). We observed no main effect of Agent, $\chi^2(2) = 2.084$, $p = 0.353$, and no significant Action by Agent interaction, $\chi^2(4) = 2.554$, $p = 0.635$.

For RT, we removed incorrect answers (3.97%) from the final dataset. We observed a main effect of Agent, $F(2139) = 24.526$, $p < 0.001$, $\eta p2 = 0.268$, a main effect of Action, $F(2140) = 48.307$, $p < 0.001$, $\eta p2 = 0.409$, and a significant Action by Agent interaction, $F(4277) = 3.311$, $p = 0.011$, $\eta p2 = 0.049$. The latter suggested that participants were faster in detecting the triangle looking up|down (0.693 ± 0.014 s) compared to looking at the graspable objects (0.738 ± 0.017 s; $p < 0.001$, |d|= 0.439, $BF_{10} = 54.749$) and to look at the table (0.785 ± 0.016 s; $p < 0.001$, |d|= 1.086, $BF_{10} = 5.479e + 10$). Participants were faster in detecting the triangle looking at the graspable objects compared to the empty table ($p < 0.001$, |d|= 0.614, $BF_{10} = 7.686e + 03$). Participants were slower in detecting the human looking at the table (0.757 ± 0.016 s) compared to the human looking up|down (0.678 ± 0.014 s; $p < 0.001$, |d|= 0.791, $BF_{10} = 2.267e + 06$) and at the graspable objects (0.686 ± 0.017 s; $p < 0.001$, |d|= 0.785, $BF_{10} = 1.833e + 06$). Participants were also slower in detecting the robot looking at the table (0.768 ± 0.017 s) compared looking up|down (0.690 ± 0.014 s; $p < 0.001$, |d|= 0.872, $BF_{10} = 3.466e + 07$) and the graspable objects (0.710 ± 0.016 s; $p < 0.001$, |d|= 0.635, $BF_{10} = 1.426e + 04$).

Furthermore, participants were slower in detecting when the triangle looked at the graspable objects compared to humans ($p < 0.001$, |d|= 0.735, $BF_{10} = 3.563e + 05$) and robots ($p = 0.002$, |d|= 0.345, $BF_{10} = 6.110$). The latter emerged as a trend on aggregated data, but still did not reach our predefined threshold of significance ($p = 0.088$). Finally, we observed faster RTs when humans looked at the table compared to the triangle ($p = 0.004$, |d|= 0.359, $BF_{10} = 8.374$). This difference revealed to be a tendency on aggregated data ($p = 0.062$).

No other Bonferroni-corrected *p*-values were lower than 0.05 for both multiple comparisons computed on the estimates of the simplified MLM and multiple comparisons using pairwise t-tests on aggregated data ($p > 0.091$, |d|< 0.332, $BF_{10} < 4.640$).

7.2 Discussion Experiment 5

Participants were faster to attribute a mental state to humans compared to the triangle. Although we observed only tendencies for robots to be faster than the triangle when gazing at the graspable object and for humans to be faster than the triangle when gazing at the table, these results are in line with Experiment 2 findings.

It is of note that responses to humans and robots looking at the graspable objects did not differ despite the visual instructions. This may appear to contradict our previous study, where we showed that participants are faster to recognise what a human compared to a robot agent is looking at (Experiment 3 in [18]). However, in Experiment 5 of the present study, the motor affordances evoked by the graspable object, or the participants' attention biased towards the side where the graspable object appeared in each trial, may explain the current results. Interestingly, this bias was present for agents with the capacity to move (robots and humans have arms) and to see (humans and robots have eyes). Such results rule out that these findings were only due to a mere attentional bias towards a hemispace of the screen. Moreover, only RTs for humans were faster than RTs for the triangle when they looked at the graspable object and tended to be faster when looking at the table. This suggests that the agent's capacity to see and form a visual representation of the surrounding (through eyes for humans and computer vision for robots) may have affected action identification differently for humans and robots.

Nonetheless, asking our participants to adopt a visual strategy may have provided two clear responses (the object and the table). Contrary, asking participants to focus on the motor

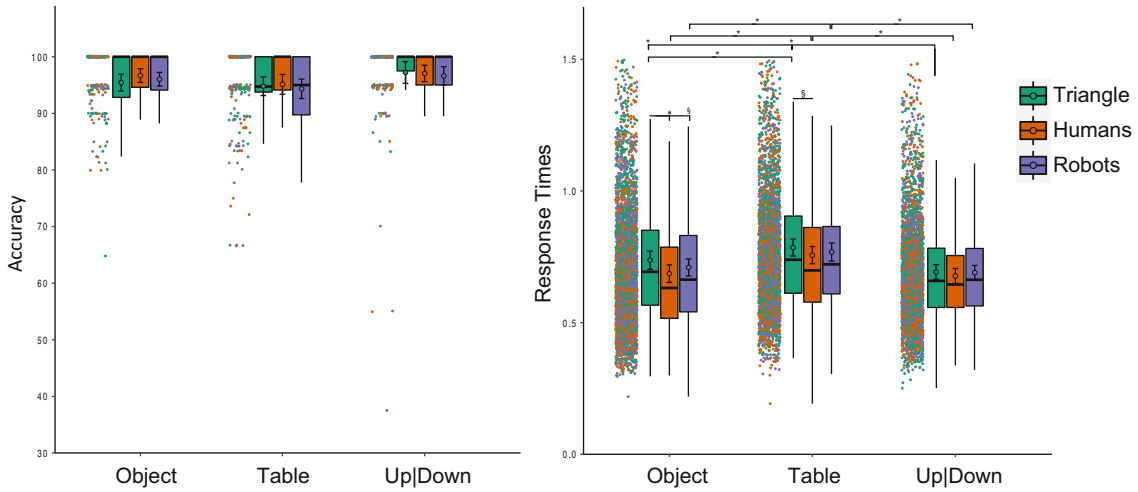

Fig. 6 Results of experiment 5, *Note.* Participants' performance in Experiment 5. The labels "Object" and "Table" indicate the conditions where participants detected the observed agent looking at the graspable object and the empty table, respectively. See Fig. 2 for a detailed explanation of our data visualisation approach

aspect (doing vs doing nothing; Experiment 4) may not have favoured the representation of two clear alternatives. Hence, we decided to perform one final experiment and change the instructions to favour a more evident non-motor alternative to the grasping condition.

8 Experiment 6

Participants (n = 81) were instructed that each agent could look towards the graspable object to grasp it ("is going to grasp the object"), towards the empty table and think about something else ("is thinking about something"), look up or look down ("is looking up or down"). Participants were asked to indicate what the agent was going to do (i.e., the agent is going to grasp the object, to think about something, or is looking up or down). Hence, we aimed to replicate the findings of Experiment 5 by providing a clearer alternative to a motor intention than we did in Experiment 4.

We removed trials with RTs deemed too fast or too slow (9.62%). Then, trials with RTs falling above or below 2.5SD of the overall mean within each block of each participant were removed (1.92%). Five participants with a performance < 65% were removed. Three participants with a performance above 2.5SD of the overall mean across conditions of the remaining participants were excluded. Despite this data management approach, three participants had 0% accuracy when the triangle looked either to the graspable object or at the empty table, and one participant had 0% accuracy when the robot looked at the empty table, suggesting a misunderstanding of the task. These participants were removed from the final sample (n = 69).

8.1 Results

We analysed performance measures (see Fig. 7) with Agent (human, robot, triangle) and Action (to grasp, to do nothing, up|down) as within-subject fixed effects of an MLM. In the case of a two-way Agent by Action interaction, we performed eighteen multiple paired comparisons of interest as indicated in Experiment 1.

For accuracy, we observed a main effect of Action, $\chi^2(2) = 17.421$, $p < 0.001$. This was not supported by the confirmatory ANOVA on aggregated data ($F = 2.832$, $p = 0.062$, $\eta p2 = 0.040$). We observed no main effect of Agent, $\chi^2(2) = 0.871$, $p = 0.647$, and no significant Action by Agent interaction, $\chi^2(4) = 1.738$, $p = 0.784$.

For RT, we removed incorrect answers (4.91%) from the final dataset. We observed a main effect of Action, $F(2136) = 43.804$, $p < 0.001$, $\eta p2 = 0.395$, a main effect of Agent, $F(2135) = 11.370$, $p < 0.001$, $\eta p2 = 0.153$, and a non-significant Action by Agent interaction, $F(4268) = 2.106$, $p = 0.080$, $\eta p2 = 0.032$.

The main effect of Action revealed that participants were slower in attributing the intention to think (0.776 ± 0.018 s) compared to the intention to grasp (0.701 ± 0.018 s; $p < 0.001$, $|d| = 1.052$, $BF_{10} = 8.232e + 09$) or look up|down (0.694 ± 0.016 s; $p < 0.001$, $|d| = 1.093$, $BF_{10} = 3.189e + 10$). We observed no difference between attributing the intention to grasp and to look up|down ($p = 1.000$, $|d| = 0.072$, $BF_{10} = 0.157$). This result may again confirm that displaying only one object may have captured participants attention and facilitated the processing of the gaze directed towards a graspable object.

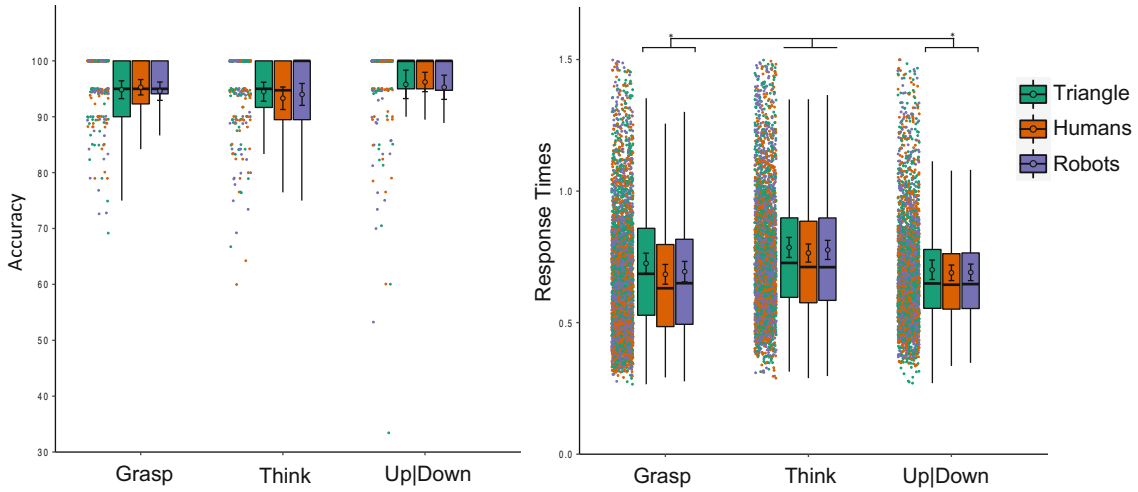

Fig. 7 Results of experiment 6, *Note.* Participants' performance in Experiment 6. No interaction between the agents and the observed action was observed for the accuracy measure. For consistency across the figures of all experiments, we display all experimental conditions. For RT, we display the main effect of Action. We invite the reader to refer to the main text for the main effect of Agent. The labels "Grasp" and "Think" indicate the conditions where participants attributed to the observed agent the intention to grasp and think, respectively. See Fig. 2 for a detailed explanation of our data visualisation approach

The main effect of Agent revealed that participants were slower in discriminating the gaze behaviour of the triangle (0.737 ± 0.018 s) compared to humans (0.713 ± 0.016 s; $p < 0.001$, |dl= 0.509, $BF_{10} = 279.605$) and robots (0.720 ± 0.016 s; $p = 0.006$, |dl= 0.341, $BF_{10} = 5.084$). We observed no difference between humans and robots ($p = 0.389$, |dl= 0.173, $BF_{10} = 0.352$).

A closer inspection to the nine experimental conditions suggests that these differences were mainly driven by a faster RT when the humans and robots gazed towards the graspable objects. Indeed, exploratory Bonferroni-corrected comparisons suggested responses to the triangle gaze were slower (0.725 ± 0.019 s) than responses to humans (0.684 ± 0.019 s; MLM $p < 0.001$, Confirmatory ANOVA $p = 0.002$, |dl= 0.495, $BF_{10} = 193.469$) and robots (0.694 ± 0.019 s; MLM $p = 0.007$, Confirmatory ANOVA $p = 0.059$, |dl= 0.367, $BF_{10} = 8.813$).

8.2 Discussion Experiment 6

Although the numerical trend was similar to the findings in Experiment 2, the change in the instructions did not provide a clearer alternative to a motor intention as we expected. In other words, we did not observe an interaction, and exploratory analyses showed not differences across agents when they looked towards the empty table. The fact that the overall RT for robots and humans did not differ (after they differed in Experiment 4) may reflect fluctuations in the data unrelated to people's general ability to attribute intentions to humans and robots (i.e., up|down condition). Similarly, RTs for robots resulted to be generally faster than the triangle (in Experiment 4 did not differ) because participants were faster in attributing to robots the intention to grasp (see Figs. 5 and 7 and Tables S3 and S4 listing all multiple comparisons for Experiments 4 and 6).

When considering the findings from Experiments 4, 5, and 6 collectively, it seems reasonable to conclude that participants used a visual strategy to solve Experiment 2. This challenges the assumption that participants were using a motor or visuo-motor strategy to attribute intentions to others. Given the high degree of concordance across the results reported in Experiments 2, 4, 5, and 6, if participants used a visual-only strategy, then the effect sizes and RTs differences computed by comparing an agent gazing at the graspable objects and towards the table should not differ across the experiments and should be similar to Experiments 5 where task instructions suggested the use of a visual-only strategy. In contrast, if a visuo-motor strategy was used, then effect sizes and RTs differences in Experiments 2, 4, and 6 should be similar to each other and differ from Experiment 5. Table 2 shows the comparison between the two highlighted conditions across the four experiments by listing several effect sizes (unstandardised β of the MLM model, the Cohen'd and BF10 computed on mean-aggregated data), and the RT difference between the two conditions based on mean- and median-aggregated data.

We note that all indexes for the triangle in Experiment 2 are the lowest, with mean and median RT differences similar to Experiment 5. Moreover, we observed that all indexes for humans and robots in Experiment 5 are lower than those in

Table 2 Comparisons between looking towards a graspable object or in the opposite direction for each observed agent

Agent	Experiment	Task instructions: "The Agent is…"	MLM	Mean-aggregated data		Average (in ms) of individual differences between conditions based on	
			|β|	Cohen's |d|	Bayes Factor	Mean	Median
Triangle	Exp.5	*"…looking at the object"* vs *"…looking at the empty table"*	0.047	0.614	7.686E + 03	48	42
	Exp.2	*"…looking at the object to grasp"* vs *"looking at the empty table"*	0.043	0.475	782.991	44	41
	Exp.4	*"…going to grasp the object"* vs *"going to do nothing"*	0.060	0.801	5.113E + 06	61	71
	Exp.6	*"…going to grasp the object"* vs *"thinking about something"*	0.062	0.727	1.803E + 05	61	65
Human	Exp.5	*"…looking at the object"* vs *"…looking at the empty table"*	0.069	0.785	1.833E + 06	70	58
	Exp.2	*"…looking at the object to grasp"* vs *"looking at the empty table"*	0.074	0.822	6.064E + 08	72	75
	Exp.4	*"…going to grasp the object"* vs *"going to do nothing"*	0.071	0.945	7.946E + 08	70	68
	Exp.6	*"…going to grasp the object"* vs *"thinking about something"*	0.083	0.841	7.203E + 06	81	73
Robot	Exp.5	*"…looking at the object"* vs *"…looking at the empty table"*	0.058	0.635	1.426E + 04	58	58
	Exp.2	*"…looking at the object to grasp"* vs *"looking at the empty table"*	0.071	0.797	2.031E + 08	70	75
	Exp.4	*"…going to grasp the object"* vs *"going to do nothing"*	0.075	0.907	2.103E + 08	74	68
	Exp.6	*"…going to grasp the object"* vs *"thinking about something"*	0.085	0.822	3.883E + 06	83	73

Comparisons between conditions where the agents looked towards the graspable objects or away from them. We present Experiment 5 at the top as it was the only experiment with instructions not evoking a motor action. We report the module of the unstandardised β derived from the MLM. We also present the module of the Cohen's d and the BF10 based on individual mean-aggregated data. The unstandardised β and the Cohen's d had all the same sign. Furthermore, we report the average (in milliseconds) of the individual differences between the two conditions based on mean- and median-aggregated data

the other experiments (except for the mean-based difference of humans in Experiment 4). These numerical trends suggest that participants were faster when they observed agents capable of grasping objects looking at graspable objects than looking at the table, but only when a motor instruction was given. These numerical observations also suggest that task instructions may have affected the strategy participants adopted. The fact that the indexes for humans and robots in Experiment 4 and 6 are similar to the indexes in Experiment 2 and higher than indexes in Experiment 5 (while the triangle's indexes in Experiment 5 are numerically similar to Experiment 2) may indicate that participants adopted a visuo-motor strategy for humans and robots only. Contrary, participants may have adopted a visual-only strategy for the triangle. However, future studies should try to disentangle

the role of visual and visuo-motor strategies when observing other people and agents gazing towards graspable objects.

8.3 Summary of all findings

Experiments 1 and 3 showed that people are faster to infer human than non-humanlike agents' intentions and suggested that humanoid robots may be perceived as part human, part non-human. Experiment 2 suggested that contextual information may facilitate recognising human-like agents' actions but not the behaviour of non-human-like agents. Experiments 4, 5, and 6 corroborated Experiment 2 and further suggested that interpreting the intention of an agent gazing toward an object may require both visual- and motor-related processes.

9 General Discussion

We investigated the ability to infer motor and communicative intentions in humans, robots, and a triangle-shaped object. Previous studies suggested that processing movements of non-human entities and attributing them human-like qualities are affected by their visual appearance [14, 43], kinematic behaviour [44], and prior assumptions [9, 45]. In all experiments, the observed kinematics did not differ across agents. Moreover, the three agents were labelled as 'human', 'robot', and 'triangle-shaped object' and described as 'agents' who could look towards different objects to perform specific actions. In particular, they could look towards an object to grasp it, look opposite to the object to self-describe (Experiments 1 and 3) or for other non-communicative purposes (Experiments 2, 4, 5, 6), and look up or down as a control condition. Such an approach ensured that we could investigate the human ability to understand the behaviour of different agents varying in their visual appearance.

We observed that when both the graspable objects and text bubble were displayed (Experiments 1 and 3), participants were faster in attributing an intention to humans compared to the robots and the triangle. Such results expand previous findings [18] and suggest that robots are perceived differently from humans when people have to infer their intentions. On the contrary, participants interpreted human and robotic gaze faster than the triangle directional changes only when the graspable object was presented alone on the table (Experiments 2, 4, 5, and 6). These findings may indicate that people can equally predict the subsequent behaviour of humans and robots from their gaze when one object prompt the location where they more likely may interact.

Results cannot be explained by participants changing the allocation of their attention and do not reflect a mere congruency effect. That is, the object on one side of the screen may have cued participants to attend and expect a gaze towards that location, and when such gaze was observed, this was considered a valid or congruent behaviour. While the graspable objects presented in isolation may have biased the attention towards one side of the screen, the faster RTs of humans and robots compared to the triangle suggest that the graspable objects may have biased not only participants' attention but may also have generated an expectation about the potential actions the agent could perform. An alternative account may suggest that the perceived objects automatically evoked affordances [40]. Thus, participants' performance may have been biased by analysing the evoked affordances rather than processing what the agent would do. However, the graspable objects may have evoked the same affordances when the triangle was the observed agent. The fact that we observed differences between human-like and non-human-like agents suggests the possibility that both the affordances evoked by the graspable objects and the capacity of the agent to act have together influenced participants' responsiveness. Hence, participants may have matched the motor possibilities of the observed agent with the evoked object's affordances while being biased towards one part of the space. It is also important to note that we observed faster RTs for humans than the triangle when the agent in question looked at the table opposite to the graspable objects (Experiments 2 and 5). Thus, participants appeared to have a behavioural advantage in recognising the goals of agents capable of acting and seeing (i.e., having a mental representation of the world). This may indicate that participants actively considered the agent's physical and mental capabilities (acting and having a representation of the table). The results across all the experiments also rule out the possibility that findings from Experiments 1 and 3 were due to differences between robot and human stimuli or that robot's stimuli may have affected human performance negatively [46].

For all these reasons, it seems evident that an integrated account is necessary to explain our results. The following sections will first discuss how our findings cannot be explained using non-inferential accounts and how results align with current theories on inferring intentions through action observation. We will also provide critical theoretical considerations and practical implications for future studies to unravel the cognitive and neural underpinnings of human–robot interactions.

9.1 Non-mentalistic Accounts do not Explain Current Results

Experiments 1 and 3 presented two objects simultaneously, and participants may have spread their attention to both sides of the screen. Consequently, the task may have been perceived more demanding than the other experiments. However, task complexity would predict a general decrease in performance with no RT differences across agents gazing

towards the graspable objects or the text bubble. Instead, we observed no differences across agents only in the non-goal directed actions. However, it may be still argued that Experiments 2, 4, 5, and 6 were more straightforward as they displayed only the graspable objects. Again, this would predict no differences in RTs across agents in all conditions. This is not supported by the observed differences between the triangle and human-like agents (humans and robots) when gazing at graspable objects.

Nevertheless, participants may not have represented others' actions in Experiments 1 and 3. Participants may have categorised what the agent was looking at despite instructions asking participants to guess what the agent would do. This explanation does not justify the different statistical results obtained across Experiments 4–5–6 and the observed differences across agents. Specifically, if instructions had no effect, it is difficult to explain why humans and the triangle differed when gazing at the table in Experiments 2 and 5 and not in Experiments 4 and 6. Therefore, given such differences across experiments, agents, and goal and non-goal directed actions, non-social and non-mentalistic strategies cannot explain our results unless considering the (human) nature of the observed agent.

9.2 Social Attention to Gaze Cues

We usually tend to follow others' gaze even when doing so is uninformative [47, 48]. From such findings, it follows that participants might infer others' attention from their gaze without engaging in other cognitive processes. In other words, participants could solve the tasks by detecting where the agent's attention was directed (i.e., towards the graspable object, the text bubble or opposite to the graspable object, up or down). A social attention account would then predict faster reaction times for humans than the triangle in all experimental conditions. However, we did not observe this advantage in non-goal-directed actions (up|down gaze, and when the triangle and humans looked away from the graspable objects in Experiments 4 and 6; see Table S3 and S4).

Another account may suggest that the labels 'human', 'robot', and 'triangle' may have favoured the perception of all non-human agents as less intentional [49]. If this were the case, we should have expected differences between humans and robots when looking at the graspable objects in Experiments 2–4–5–6 and when performing non-goal-directed gazes. We did not observe such results. However, it may be proposed that the triangle did not engage with participants as it has no eyes. Results may thus reveal a lack of social engagement [48] of the triangle compared to humans and robots who directly stared at participants. However, we exclude this possibility as participants had comparable RTs across all agents in the control condition (up|down gaze). Moreover, we found no differences between the triangle and the robots (which had eyes) in the look-away conditions in Experiments 2–4–5–6. In a similar vein, results from Experiments 1 and 3 cannot be explained by a social attention account alone, as we did not find an advantage (faster RTs) in detecting humans compared to other agents' non-goal-directed actions. Finally, if our description may have helped the perception of robots as intentional agents, we should have observed no difference between humans and robots when gazing towards the graspable objects in Experiments 1 and 3.

Given the current findings, it is clear that the non-human nature and non-human-like appearance of the observed agents modulated the readiness to predict their actions following an observed gaze.

9.3 Evidence for an Integrated Account

Across all experiments, participants first saw the environment (table, objects, and text bubble for Experiments 1 and 3), then the agent, and only after a short interval the agent's gaze. Thus, participants had time to process the objects on the table and their affordances, social information like the (non-) human nature of the agent and its visual shape, and motor information from the agents' gaze separately. Hence, it is possible that the observer's motor and visual system analysing the agent's action and bodily form interacted during task completion. In particular, participants may have integrated what the agent was looking at, its ability to act based on its human-like shape, and its mental content based on its human and non-human nature.

In Experiment 2, participants were faster when humans and robots gazed towards the graspable objects compared to when the triangle directed attention to these graspable objects (a similar trend was observed in Experiments 4 and 6). This suggests that the visual form of the agent modulated the ability to anticipate what the agent was going to do, and that their gaze was perceived as a plausible goal-directed action (i.e., humans and humanoid robots can grasp an object while a triangle without effectors cannot). Furthermore, the slow RT when the agents looked opposite to the graspable objects in Experiments 2, 4, 5, and 6 may indicate that looking away from a graspable object was unexpected (i.e., people may have directed their attention more on one side of the screen). Importantly, we found that when task instructions may have emphasised a visual strategy and focused on what the agent was looking at (Experiments 2 and 5), participants had faster RTs for humans than the triangle. This confirms that it is more challenging to indicate what a non-human rather than a human agent sees. However, we also observed that RTs in Experiment 5 were slightly faster for robots than the triangle when looking at the graspable objects. This result does not contradict the findings from our previous study (Experiment 3 [18]) where we observed no RT differences between robots and the triangle. However, in that experiment, the graspable

objects and text bubble were displayed simultaneously and participants indicated what the agents were looking at (i.e., at the object, at the text bubble). Therefore, this apparently contradictory results may indicate an interaction between attentional processes, the analysis of the body form of the agent, and the processing of the affordances evoked by the graspable object. In Experiment 5, participants' attention was biased towards the graspable object, and a gaze towards it may have been perceived more plausible for agents possessing hand effectors (i.e., humans and robots) than the triangle. Hence, the faster RT of robots compared to the triangle when gazing towards the graspable objects suggests that participants might have predicted such gaze [50] given the robots' psychical capabilities and the fact that only one object was shown. However, when participant's attention is not biased, such an advantage is not present (see Experiment 3 in [18]).

Overall, we observed differences across agents in experimental conditions evoking goal-directed behaviours (gazing at the graspable objects and text bubble in Experiments 1–3; gazing at the graspable objects in Experiments 2–4–6), mental representations of the surrounding (looking at the table in Experiments 2 and 5). On the contrary, we observed no differences for non-goal-directed actions (looking at the table described as a non-motor and non-visual action in Experiments 4 and 6; upldown gaze in all Experiments). These results suggest that participants were not merely detecting the focus of the agent's attention but were also inferring the motivation guiding the agent's gaze (grasp, speak, look). Hence, our experiments as a whole confirm that differences in understanding others' gazes are more likely to emerge during the observation of goal-directed rather than non-goal-directed actions and when actively thinking about the agent's mental state [18].

For all these reasons, it is plausible to consider that the processes analysing the agents' body form and the tendency to attribute high-level social skills to humans rather than non-human agents contributed to participants' ability to understand their gaze. In particular, the findings reported in this study suggest that processing an agent's bodily form and considering its motor abilities, together with the goal-directedness of the observed behaviour, may contribute to how observers explicitly predict what the agent is doing from their gaze. As such, this work extends current models of gaze perception that establish a role for gaze motion and mentalising processes (thought to be supported by brain regions including the superior temporal sulcus, medial prefrontal cortex), attentional mechanisms (supported by parietal lobe), and top-down processes (i.e., adopting an intentional stance or not; [6, 7, 51]) to infer others' mental states.

9.4 Is Predicting Action from Gaze Equivalent to Intention Reading?

We can identify three processes supporting the ability to infer others' intentions and goals from action observation. When observing others behaviour, we may identify what the person is doing (e.g., grasping), how the action is performed (e.g., using a whole or precision grasp), and why (e.g., to eat or to give). Across the six experiments, participants had to predict the intentions the observed gaze served (e.g., gazing rightwards to grasp). Specifically, we asked participants to observe an agent who averted their gaze ('What') by looking towards four directions ('How': right, left, up, down) with different final goals ('Why': grasping, speaking, looking, thinking, doing nothing).

A motor action is defined as a sequence of motor acts directed towards a distal goal (e.g., reaching for a piece of food, grasping it, holding it, and bringing it to the mouth). Hence, the look-to-grasp and look-to-speak actions showed in our experiments are comparable to the typical grasp-to-eat and grasp-to-drink action chain. This interpretation may be criticised as grasping and talking may be considered the 'what' an agent is doing. It is important to note that while others gaze contributes to action prediction abilities [52, 53], action observation studies typically focus the observer's attention on the actor's upper limb without displaying the actor's gaze and face [54–58]. Studying the temporal order of action is a relatively new area of inquiry [59], which one day may refine the criteria defining the 'what' and the 'why' of the observed action.

Some critiques have also been raised to such a hierarchical view of intention reading from action observation (how < what < why), suggesting that the question participants should answer to investigate the hidden 'why' of an observed behaviour should have a good degree of abstraction to generalise across different actions goals [60]. However, we note that the number of hidden states behind an observed act may be infinite. For example, a person may gaze out of their car window to locate a spot to park their car to go to the theatre to watch the new show so that they can then write a review on it to be published in the school newspaper to impress the headteacher (and so on). In this example, it is hard to clearly define which end-state has a good level of abstraction or should be considered the ultimate goal of a sequence of actions. In real life, reading others' intentions relies on inferring (maybe erroneously) their hidden states using a limited set of behavioural and contextual information available at a given time. In a similar way, we used instructions to provide an adequate context and facilitate the different levels of representation of the observed behaviour [61].

To summarise, we believe that grasping and speaking, as displayed in our tasks, can be considered as immediate (proximal) intentions within the gaze-to-grasp and gaze-to-speak action chain.

9.5 Limitations of the current study

In this series of experiments, we compared human and human-like agents with non-human-like agents. It is important to note that both humans and human-like agents differed in their biological nature (humans are living entities, while robots are not), visual appearance (Nao and Baxter had different sizes and textures compared to humans), as well as behavioural repertoire (e.g., Nao has a face with eyes while Baxter has a display, both Nao and Baxter hand effectors differ from a human hand). On the contrary, the non-human-like agent (the triangle) was not only a non-biological agent but also did not possess any grasping capabilities. However, while Experiments 1 and 3 suggest that the motor capabilities of an agent may not be crucial to predicting the action of a robot with a human-like shape (RTs for humans were faster than the robots and the triangle), Experiments 2, 4, 5, and 6 did not test how quickly participants can predict the behaviour of a mechanical robot with grasping capabilities but non-human-like appearance. Further studies will be required to assess the interaction between human-like appearance and motor capabilities in predicting the behaviour of human-like and non-human-like robots from gaze observation.

10 Conclusions

We observed that an agent's visual body-form (human-like vs non-human-like) and biological nature (human vs non-human) may differently affect the processing of high-level social behaviours from gaze observation.

Results expand current models of gaze perception and cannot be explained by non-motor and non-social cognitive mechanisms. Instead, mindreading from action observation likely integrates the observed agent's intentional nature with perceptual, motor, and mentalising processes. Moreover, our experimental designs and stimuli have shown that biasing participants' attention may facilitate predicting what a robot might do. In other words, visual information that cue where a robot may act and may help an observer or collaborator understand what the robot will do within social and non-social contexts. An applied and testable outcome of our results is that visual and body cues can be used to facilitate the intuitive interpretation of humanoid robot behaviour [62]. For example, this might be achieved by positioning the robot's body in a way that the grasping hand is closer to the object-to-be-grasped. Contrary, having both effectors equally close to two objects may not help the observer predict what the robot will do next.

Overall, these findings have important theoretical, methodological, and applied implications for human–human and human–robot social cognition and the development of successful and trustworthy human–robot collaboration.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12369-022-00962-2.

Acknowledgements We thank Loron Hill for the interesting discussions on preliminary results.

Author's Contribution ET conceptualized the study. ET programmed the experiments and collected the data. ET and MS analyzed the data. ET, ESC, MS, NC interpreted the results and wrote the manuscript.

Funding While this study was not directly funded by any grants, ESC's contributions were supported in part by the European Research Council (ERC) under the EU Horizon 2020 research and innovation program (grant agreement 677270) and the Leverhulme Trust (PLP-2018-152), and NC's contributions were supported by a Macquarie University Research Fellowship.

Data Availability The datasets and scripts of the current study are available in the OSF repository, https://osf.io/bd6h3/.

Declarations

Conflict of interest ESC is member of the editorial board of the journal.

References

1. Baron-cohen S, Baron-cohen S, Centre AR, Centre AR (2005) The empathizing system: a revision of the 1994 model of the mindreading system. Mind 1–44
2. Tomasello M (2010) Origins of human communication. MIT press, Cambridge
3. Yu C, Smith LB (2017) Multiple sensory-motor pathways lead to coordinated visual attention. Cogn Sci 41:5–31. https://doi.org/10.1111/cogs.12366

4. Caruana N, Inkley C, Nalepka P et al (2021) Gaze facilitates responsivity during hand coordinated joint attention. Sci Rep 11:21037. https://doi.org/10.1038/s41598-021-00476-3
5. Emery NJ (2000) The eyes have it: the neuroethology, function and evolution of social gaze. Neurosci Biobehav Rev 24:581–604. https://doi.org/10.1016/S0149-7634(00)00025-7
6. Teufel C, Fletcher PC, Davis G (2010) Seeing other minds: attributed mental states influence perception. Trends Cogn Sci 14:376–382. https://doi.org/10.1016/j.tics.2010.05.005
7. Wykowska A, Wiese E, Prosser A, Müller HJ (2014) Beliefs about the minds of others influence how we process sensory information. PLoS One 9:e94339. https://doi.org/10.1371/journal.pone.0094339
8. Kühn S, Brick TR, Müller BCN, Gallinat J (2014) Is this car looking at you? How anthropomorphism predicts fusiform face area activation when seeing cars. PLoS One 9:1–14. https://doi.org/10.1371/journal.pone.0113885
9. Stanley J, Gowen E, Miall RC (2007) Effects of agency on movement interference during observation of a moving dot stimulus. J Exp Psychol Hum Percept Perform 33:915–926. https://doi.org/10.1037/0096-1523.33.4.915
10. Klapper A, Ramsey R, Wigboldus D, Cross ES (2014) The control of automatic imitation based on bottom–up and top-down cues to animacy: insights from brain and behavior. J Cogn Neurosci 26:2503–2513. https://doi.org/10.1162/jocn_a_00651
11. Caruana N, Spirou D, Brock J (2017) Human agency beliefs influence behaviour during virtual social interactions. PeerJ 5:e3819. https://doi.org/10.7717/peerj.3819
12. Caruana N, McArthur G (2019) The mind minds minds: the effect of intentional stance on the neural encoding of joint attention. Cogn Affect Behav Neurosci 19:1479–1491. https://doi.org/10.3758/s13415-019-00734-y
13. Ramsey R (2018) Neural integration in body perception. J Cogn Neurosci 30:1442–1451. https://doi.org/10.1162/jocn_a_01299
14. Morales-Bader D, Castillo RD, Olivares C, Miño F (2020) How do object shape, semantic cues, and apparent velocity affect the attribution of intentionality to figures with different types of movements? Front Psychol 11:1–14. https://doi.org/10.3389/fpsyg.2020.00935
15. Quadflieg S, Mason MF, Macrae CN (2004) The owl and the pussycat: gaze cues and visuospatial orienting. Psychon Bull Rev 11:826–831. https://doi.org/10.3758/BF03196708
16. Downing P, Dodds C, Bray D (2004) Why does the gaze of others direct visual attention? Vis cogn 11:71–79. https://doi.org/10.1080/13506280344000220
17. Baron-Cohen S (1994) The mindreading system: new directions for research. Curr Psychol Cogn 13:724–750
18. Tidoni E, Holle H, Scandola M et al (2022) Human but not robotic gaze facilitates action prediction. iScience 25:104462. https://doi.org/10.1016/j.isci.2022.104462
19. Fitter NT, Kuchenbecker KJ (2016) Designing and assessing expressive open-source faces for the baxter robot. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp 340–350
20. Palan S, Schitter C (2018) Prolific.ac—A subject pool for online experiments. J Behav Exp Financ 17:22–27. https://doi.org/10.1016/j.jbef.2017.12.004
21. Faul F, Erdfelder E, Lang A-G, Buchner A (2007) G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 39:175–191. https://doi.org/10.3758/BF03193146
22. Nomura T, Kanda T, Suzuki T, Kato K (2008) Prediction of human behavior in human–robot interaction using psychological scales for anxiety and negative attitudes toward robots. IEEE Trans Robot 24:442–451. https://doi.org/10.1109/TRO.2007.914004
23. Bartneck C, Kulić D, Croft E, Zoghbi S (2009) Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Int J Soc Robot 1:71–81. https://doi.org/10.1007/s12369-008-0001-3
24. Schubert TW, Otten S (2002) Overlap of self, ingroup, and outgroup: pictorial measures of self-categorization. Self Identity 1:353–376. https://doi.org/10.1080/152988602760328012
25. Kleiner M, Brainard D, Pelli D (2007) What's new in psychtoolbox-3. Perception 36:1–16
26. Peirce J, Gray JR, Simpson S et al (2019) PsychoPy2: experiments in behavior made easy. Behav Res Methods 51:195–203. https://doi.org/10.3758/s13428-018-01193-y
27. Tamir DI, Thornton MA, Contreras JM, Mitchell JP (2016) Neural evidence that three dimensions organize mental state representation: rationality, social impact, and valence. Proc Natl Acad Sci 113:194–199. https://doi.org/10.1073/pnas.1511905112
28. Shiffrar M, Freyd JJ (1990) Apparent motion of the human body. Psychol Sci 1:257–264. https://doi.org/10.1111/j.1467-9280.1990.tb00210.x
29. Schenke KC, Wyer NA, Bach P (2016) The things you do: internal models of others' expected behaviour guide action observation. PLoS One 11:1–22. https://doi.org/10.1371/journal.pone.0158910
30. Foundation RCTR (2017) R: A Language and Environment for Statistical Computing. Study R https://www.R-project.org
31. Bates D, Mächler M, Bolker BM, Walker SC (2015) Fitting linear mixed-effects models using lme4. J Stat Softw. https://doi.org/10.18637/jss.v067.i01
32. Ben-Shachar M, Lüdecke D, Makowski D (2020) Effectsize: estimation of effect size indices and standardized parameters. J Open Source Softw 5:2815. https://doi.org/10.21105/joss.02815
33. Lüdecke D, Ben-Shachar M, Patil I et al (2021) Performance: an R package for assessment, comparison and testing of statistical models. J Open Source Softw 6:3139. https://doi.org/10.21105/joss.03139
34. Kamil B (2016) MuMIn: multi-model inference. R Packag Version 1:1–15
35. Lenth R (2019) Emmeans: estimated marginal means. In: R Packag. Version 1.4.2
36. Raftery AE (1995) Bayesian model selection in social research. Sociol Methodol 25:111. https://doi.org/10.2307/271063
37. Allen M, Poggiali D, Whitaker K et al (2019) Raincloud plots: a multi-platform tool for robust data visualization. Wellcome Open Res 4:63. https://doi.org/10.12688/wellcomeopenres.15191.1
38. Handy TC, Grafton ST, Shroff NM et al (2003) Graspable objects grab attention when the potential for action is recognized. Nat Neurosci 6:421–427. https://doi.org/10.1038/nn1031
39. Franca M, Turella L, Canto R et al (2012) Corticospinal facilitation during observation of graspable objects: a transcranial magnetic stimulation study. PLoS One. https://doi.org/10.1371/journal.pone.0049025
40. Bach P, Nicholson T, Hudsons M (2014) The affordance-matching hypothesis: how objects guide action understanding and prediction. Front Hum Neurosci 8:1–13. https://doi.org/10.3389/fnhum.2014.00254
41. Bukowski H, Hietanen JK, Samson D (2015) From gaze cueing to perspective taking: revisiting the claim that we automatically compute where or what other people are looking at. Vis Cogn 23:1020–1042. https://doi.org/10.1080/13506285.2015.1132804
42. Furlanetto T, Becchio C, Samson D, Apperly I (2016) Altercentric interference in level 1 visual perspective taking reflects the ascription of mental states, not submentalizing. J Exp Psychol Hum Percept Perform 42:158–163. https://doi.org/10.1037/xhp0000138
43. Li AX, Florendo M, Miller LE, et al (2015) Robot form and motion influences social attention. ACM/IEEE Int Conf Human-Robot Interact 2015-March:43–50. https://doi.org/10.1145/2696454.2696478

44. Cross ES, Liepelt R, Antonia AF et al (2012) Robotic movement preferentially engages the action observation network. Hum Brain Mapp 33:2238–2254. https://doi.org/10.1002/hbm.21361
45. Cross ES, Ramsey R, Liepelt R et al (2016) The shaping of social perception by stimulus and knowledge cues to human animacy. Philos Trans R Soc B Biol Sci. https://doi.org/10.1098/rstb.2015.0075
46. Mandell AR, Smith M, Wiese E (2017) Mind perception in humanoid agents has negative effects on cognitive processing. Proc Hum Factors Ergon Soc 2017:1585–1589. https://doi.org/10.1177/1541931213601760
47. Driver J, Davis G, Ricciardelli P et al (1999) Gaze perception triggers reflexive visuospatial orienting. Vis Cogn 6:509–540. https://doi.org/10.1080/135062899394920
48. Kampe KKW, Frith CD, Frith U (2003) "Hey John": signals conveying communicative intention toward the self activate brain regions associated with "mentalizing", regardless of modality. J Neurosci 23:5258–5263. https://doi.org/10.1523/jneurosci.23-12-05258.2003
49. Wiese E, Wykowska A, Zwickel J, Müller HJ (2012) I see what you mean: how attentional selection is shaped by ascribing intentions to others. PLoS One 7:1–7. https://doi.org/10.1371/journal.pone.0045391
50. Joyce K, Schenke K, Bayliss A, Bach P (2016) Looking ahead: anticipatory cueing of attention to objects others will look at. Cogn Neurosci 7:74–81. https://doi.org/10.1080/17588928.2015.1053443
51. Stephenson LJ, Edwards SG, Bayliss AP (2021) From gaze perception to social cognition: the shared-attention system. Perspect Psychol Sci 16:553–576. https://doi.org/10.1177/1745691620953773
52. Ramsey R, Cross ES, de Hamilton AFC (2012) Predicting others' actions via grasp and gaze: evidence for distinct brain networks. Psychol Res 76:494–502. https://doi.org/10.1007/s00426-011-0393-9
53. Pierno AC, Becchio C, Wall MB et al (2006) When gaze turns into grasp. J Cogn Neurosci 18:2130–2137. https://doi.org/10.1162/jocn.2006.18.12.2130
54. Bianco V, Finisguerra A, Betti S et al (2020) Autistic traits differently account for context-based predictions of physical and social events. Brain Sci 10:1–20. https://doi.org/10.3390/brainsci10070418
55. Amoruso L, Finisguerra A, Urgesi C (2020) Spatial frequency tuning of motor responses reveals differential contribution of dorsal and ventral systems to action comprehension. Proc Natl Acad Sci U S A 117:13151–13161. https://doi.org/10.1073/pnas.1921512117
56. Suttrup J, Keysers C, Thioux M (2015) The role of the theory of mind network in action observation—an rTMS study. Brain Stimul 8:415–416. https://doi.org/10.1016/j.brs.2015.01.326
57. Becchio C, Manera V, Sartori L et al (2012) Grasping intentions: from thought experiments to empirical evidence. Front Hum Neurosci 6:1–6. https://doi.org/10.3389/fnhum.2012.00117
58. Errante A, Ziccarelli S, Mingolla GP, Fogassi L (2021) Decoding grip type and action goal during the observation of reaching-grasping actions: a multivariate fMRI study. Neuroimage 243:118511. https://doi.org/10.1016/j.neuroimage.2021.118511
59. Thomas RM, De Sanctis T, Gazzola V, Keysers C (2018) Where and how our brain represents the temporal structure of observed action. Neuroimage 183:677–697. https://doi.org/10.1016/j.neuroimage.2018.08.056
60. Thompson EL, Bird G, Catmur C (2019) Conceptualizing and testing action understanding. Neurosci Biobehav Rev 105:106–114. https://doi.org/10.1016/j.neubiorev.2019.08.002
61. Grafton ST, Tipper CM (2012) Decoding intention: a neuroergonomic perspective. Neuroimage 59:14–24. https://doi.org/10.1016/j.neuroimage.2011.05.064
62. Setchi R, Dehkordi MB, Khan JS (2020) Explainable robotics in human–robot interactions. Procedia Comput Sci 176:3057–3066. https://doi.org/10.1016/j.procs.2020.09.198

Michele Scandola is a cognitive neuroscientist and Assistant Professor at the University of Verona, Italy. His main research interests include the representation of the body in healthy and CNS-damaged people, the representation of space, and the impact of social, temporal or spatial conditions in decision-making and other cognitive functions. He is an expert in Bayesian Statistics and has developed specific Bayesian statistical models for studying single cases and peripersonal space. He provided evidence for the impact of spinal cord injuries on the cognitive representations concerning the body, action, space and empathic functions.

Emily S. Cross is a cognitive and social neuroscientist who directs the Social Brain in Action research laboratory, which is based jointly at the Institute of Neuroscience and Psychology at the University of Glasgow in Scotland, and the MARCS Institute for Brain, Behaviour and Development at Western Sydney University in Australia. Using intensive training procedures, functional neuroimaging, brain stimulation, and research paradigms involving dance, acrobatics and robots, she leads a team who explores questions concerning how we learn via observation, motor expertise, and social influences on human—robot interaction. She and her team are particularly interested in how prolonged experience with robots changes how we perceive and interact with embodied artificial agents at brain and behavioural levels, and how these relationships manifest across the lifespan and in different cultures.

Nathan Caruana is a cognitive neuroscientist and Senior Research Fellow at Macquarie University. His research examines the mechanisms that support positive experiences and collaboration during social interactions between humans and with artificial agents (virtual characters and robots). He has developed several interactive eye-tracking and virtual reality paradigms which interface with human neuroimaging, neurophysiology and behavioural methods. These methods have provided novel insights into the mechanisms that allow humans to perceive, evaluate and respond to social information in the context of reciprocal, dynamic and collaborative social interactions. His research is specifically focused on understanding the various perceptual (bottom-up) and psychological (top-down) factors that shape how people understand and coordinate with others, and how the role of these factors may vary across the neurodiverse human population.

Emmanuele Tidoni is a social neuroscientist and Lecturer at the University of Hull in the United Kingdom. His research focuses on how people perceive and understand others' behaviour as intentional and goal-directed. He combines behavioural and non-invasive brain stimulation methodologies to study the cognitive and neural mechanisms supporting human social cognition. His innovative paradigms combine the use of humans and artificial devices to apply and expand current models of social cognition to human and non-human agents.