International Journal of Social Robotics
https://doi.org/10.1007/s12369-023-01036-7

ORIGINAL RESEARCH

Try to See it My Way: Humans Take the Level-1 Visual Perspective of Humanoid Robot Avatars

Basil Wahn[1,2] · Leda Berio[3] · Matthias Weiss[4] · Albert Newen[3]

Accepted: 24 July 2023

Abstract
Visual perspective taking (VPT) is a fundamental process of social cognition. To date, however, only a handful of studies have investigated whether humans also take the perspective of humanoid robots. Recent findings on this topic are conflicting as one study found no evidence for level 1 VPT (i.e., which object is seen by the agent) and a different study has found evidence for level 2 VPT (i.e., how is the object seen by the agent). The latter study proposed that the human-like appearance of robots triggers VPT and that a mental capacity to perceive the environment is not required (mere-appearance hypothesis). In the present study, we tested whether the mere-appearance hypothesis is also applicable to level 1 VPT. We manipulated the appearance of a humanoid robot by either showing it with a human-like or artificial head, and its mental capacity for perception by presenting it as switched on or off. We found that all manipulations triggered VPT, showing, in contrast to earlier findings, level 1 VPT for robots. Our findings support the mere-appearance hypothesis as VPT was triggered regardless of whether the robot was switched on or off, and also show that the mere-appearance hypothesis is robust with regard to alterations of human-like appearance.

Keywords Social cognition · Visual perspective taking · Human–robot interaction · Humanoid robots

1 Introduction

Visual perspective taking (VPT), i.e. the ability to adopt others' visual perspective, is a fundamental process of social cognition, which is particularly important for communication [38]. To better classify the abilities involved in VPT, Flavell et al. [8] early on introduced the distinction between level 1 VPT (henceforth "VPT1") and level 2 VPT (henceforth "VPT2"). VPT1 involves the ability to identify *what* lies in others' line of sight (i.e., which objects are seen by another agent) whereas VPT2 involves "mentally adopting someone else's spatial point of view and understanding *how* the world is represented from this virtual perspective" [18]. That is, while VTP1 involves the ability to register which objects from your own perspective another person can (and cannot) see, VPT2 involves the ability to register how objects are seen from the perspective of another person.

Supporting this distinction made by Flavell et al. [8], there are a number of key differences between VPT1 and VPT2. VPT2 occurs later in development and can be impaired in clinical populations where VPT1 is intact [14]. Also, it was suggested that VPT2 involves egocentric mental rotations, which is not true for VPT1 [36]. Moreover, the two types of VPT are potentially differently affected by the beliefs of bystanders [6]. In particular, it was found that a bystander's task-irrelevant belief was increasing speed in reaction times during a VPT1 object tracking task but not during a VPT2 object tracking task. Finally, Martin et al. [21] found that, compared to younger adults, older adults were less influenced by others' perspective in a VPT1 task, but the reverse was true for a VPT2 task. Taken together, these findings suggest that processes involved for VPT2 are not necessarily the same processes involved in VPT1 and it is thus important to

Basil Wahn and Leda Berio shared first authorship.

✉ Basil Wahn
basil.wahn@rub.de

1 Institute of Educational Research, Ruhr University Bochum, Universitätsstraße 150, 44801 Bochum, Germany

2 Department of Neurophysiology and Pathophysiology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

3 Department of Philosophy, Ruhr-Universität Bochum, Bochum, Germany

4 Department of Business & Economics, Zeppelin University, Friedrichshafen, Germany

Published online: 30 August 2023

investigate whether findings for VPT1 generalize to VPT2 and vice versa.

One task that has been particularly influential when investigating VPT1 is the dot-matching task by Samson et al. [30]. In this task, participants were required to indicate whether a predefined number of dots match the number of dots that can be seen either from their own perspective or the perspective of a human avatar. Importantly, it was found that a human avatar's perspective interfered with participants' judgments from their own perspective (henceforth referred to as "altercentric intrusions"). Moreover, they found that when participants were required to judge from the avatar's perspective, their own perspective interfered with their judgment from the avatar's perspective (henceforth referred to as "egocentric intrusions"). While egocentric intrusions were expected, the altercentric intrusions constituted novel and compelling evidence for VPT1. Since this study, several studies have extended this work (e.g., [35, 44]). Following these results, a question that has emerged in the literature is to what extent VPT1 is spontaneous and/or automatic, i.e. to what extent it occurs rapidly and involuntarily (hence spontaneous) and to what extent it is also reflexive, stimulus-driven, and not subject to inhibition (hence automatic) (for a literature overview, see [28]). O'Grady et al. [29] argued that VPT1 should be considered a spontaneous process and supported their claim with results from a series of experiments, in which they tested factors that are predicted to affect VPT1 if it is a spontaneous process.

To date, however, VPT1 has been primarily investigated with human avatars but not with robot avatars. As robots are increasingly becoming important actors in our social world [4, 33], questions arise about the extent to which our interactions with social robots resemble that with other human agents [2]. Especially, given the importance of VPT when it comes to social interactions [38, 39], it is crucial to assess the ways in which humans can assume a robot's perspective, thereby gaining insights about the interaction between humans and social robots [34]. Several studies have assessed to what extent the altercentric intrusions described above are exclusive to the presence of human avatars, for example, it has been shown that VPT1 is not triggered by the presence of objects [27, 30, 37]. However, to what extent the presence of a robotic avatar can trigger these VPT-effects is still not well understood. This study aims at addressing this lacuna.

1.1 Robots and VPT

So far, results on whether humans take the perspective of a robot have been mixed. A study by Xiao et al. [44] adapted the design of Samson et al. [31] to test whether VPT1 occurs when a humanoid robot avatar is presented. Their study was also a lab experiment with a four times larger sample size ($N = 64$) than Samson et al. [30] to ensure sufficient statistical power. The humanoid robot avatar they used had a clear humanoid body and head shape. As a control condition, they also ran the study with human avatars. Their results suggest that robotic avatars, as opposed to human avatars, do not trigger VPT1. However, with regard to VPT2, an online study by Zhao and Malle [45] showed that humanoid robots do trigger VPT2. In their study, participants performed a number identification task, in which a number on a table is seen as a 9 from the participant's perspective and as a 6 from the robot's perspective. Given that the two perspectives are conflicting with regard to *how* the number is seen, this task can be used to infer to what extent humans take the perspective of the robot in a VPT2 task [45]. In their design, they used a wide range of robots that varied in their human-like appearance. These included robots with a humanoid body and head or with neither of these attributes (i.e., a box-like robot). Also, a human-like doll and a cat were presented. Each agent was tested with a large sample size ($N = 100$). They found VPT2 for robots with a humanoid body and head and for the doll but not for the box-like robot and cat.

Zhao and Malle [45] took the data as strong evidence that human-likeness triggers VPT as a visual association process without any need of involving an explicit attribution of a mental capacity. In particular, they claim it supports their "mere-appearance hypothesis", which is the idea that human-like appearance triggers VPT towards robots. Zhao and Malle base this hypothesis on the idea that very familiar stimulus responses are extended to new stimuli if they resemble the original [13, 32], as well as on the fact that recent evidence suggests that human-like robots are likely to trigger various socio-cognitive processes like gaze following and anthropomorphization [5, 7, 23, 40]. The central idea of the mere-appearance hypothesis is that human-likeness can trigger VPT, in a way that is unmediated by any attribution of a mental capacity enabling an agent to actually look at an object and make sense of the object. Moreover, it predicts that with increasing human-likeness of robots, the likelihood of VPT increases. In this sense, the mere-appearance hypothesis is presented as in contrast with the mind-perception hypothesis [12, 42], which predicts that perspective taking depends on attribution of human-like mind abilities (and thus partially on Theory of Mind), and also with the uncanny-valley-hypothesis, which predicts that excessive human-likeness would impair perspective taking [22].

Collectively, the findings by Xiao et al. [43] and Zhao and Malle [45] present us with some possible challenges. If Zhao and Malle are correct that human-like appearance can trigger VPT2, one would naturally wonder whether the same is true for VPT1 given that findings for VPT2 do not necessarily generalize to VPT1 (and vice versa) as noted above. On the one hand, if human-likeness can cause subjects to adapt the visual perspective of others and compute

spontaneously *how* something is seen, we might expect it to also modulate the more (ontogenetically and phylogenetic) basic skills implied in identifying what is in the line sight of others involved in VPT1 [14], and therefore to possibly also generate altercentric intrusions. On the other hand, one might think that human-likeness only becomes a relevant variable once VPT2 is required, in other words, it could be that human-likeness can impact VPT2 but does not modulate VPT1 because the ability to infer how an object is seen by others may only be relevant when interacting with other humans. Human-likeness might also only be a variable relevant with more explicit, time consuming judgements like those in Zhao and Malle's [45] task, but not when it comes to spontaneous (and hence rapid and involuntary) VPT like the one involved in the dot-matching task [30]. To address these issues, we test VPT1 altercentric and egocentric intrusions using two different human-like robots in the dot-matching task [30].

Another point of divergence between the studies by Xiao et al. [43] and Zhao and Malle [44] are the features suggesting gaze and direction. While Zhao and Malle present a large variety of robots with clearly visible features indicating looking direction (such as eyes), Xiao and colleagues present a humanoid robot only in one version, with the robot's features indicating the direction of sight not being clearly visible. Whether these features are clearly visible or not may be of high importance given that not only the human-like appearance is critical to trigger VPT according to the mere-appearance hypothesis but also the looking direction [45].

To address these points, as Xiao et al. [43], we also used the paradigm by Samson et al. [30] and present humanoid robot avatars but now with clearly visible features that indicate the direction of sight to test whether they trigger VPT1. Also, to probe the extent to which human appearance triggers the presence of VPT1, we vary the appearance of the robot. That is, we present the participants with two different kinds of humanoid robotic avatars, one with a human-like head (with a human-like visual system), and one with an artificial camera-like head (with an artificial looking visual system). This should inform us on the effect that the presence of a human-like visual system has on VPT1. In other words, is it critical for VPT1 that the direction of sight is clearly indicated with a human visual system or is VPT1 also triggered with a clearly indicated direction of sight with an artificial looking visual system? To manipulate the robot's mental capacity for perception, we present the robots as either switched on or switched off, by means of introducing them as such in the instructions, and by reminding the participants of the status displaying a green or red indicator light on the robot. This should inform us on whether VPT1 is dependent on the presence of a mental capacity enabling the ability to see an object or not, according to the mere-appearance hypothesis, perception of mental capacity should not be necessary for altercentric intrusions. As a point of note, we opted for this manipulation to vary the robot's capabilities for perception as it does not involve altering the visual system of the robot (e.g., by closing the eyes) and thus does not introduce additional confounding factors that could have alternatively explained our results (e.g., open vs. closed eyes).

If we find altercentric intrusions only when the robot is switched on but not when it is switched off, then this would speak in favor of a mind-perception account, where the mental abilities are required to trigger VPT (in this case, the ability for perception). Altercentric intrusions in both conditions, conversely, would constitute possible supporting evidence for the mere-appearance hypothesis, as the appearance of the robot alone would be sufficient for VPT1.

By varying the appearance of the robot (i.e., either showing a human-like or artificial-looking head), we test the boundaries of the mere-appearance hypothesis more directly. That is, if the shape of the head is particularly relevant for human-likeness, we would expect altercentric intrusions only for the human-like head but not for the artificial head. Conversely, if we find altercentric intrusions for the human-like as well as the artificial head, then this would indicate that other human-like elements of the robot's body may trigger VPT.

2 Methods

2.1 Participants

We collected data from 128 participants ($M = 29.23$ years, $SD = 4.00$ years, 67 female, 59 male, 2 diverse) located in the US and UK via the participant recruitment service *Prolific*. 32 participants participated in each of our between-subject conditions (On + Human Head, Off + Human Head, On + Artificial Head, Off + Artificial Head). To replicate the egocentric and altercentric intrusion effects found by Samson et al. [30], a sample size of 16 participants for each condition would have matched the sample size by Samson et al. [30]. A sample size of 16 participants per condition, however, would only be sufficient to detect large effects (Cohen's $d = 1.02$; Power = 0.80, alpha = 0.05, two-tailed independent t-test; Software used: *G*Power*). Given that we investigate with our between-subject conditions novel effects for which the effect size is unknown, we increased the sample size to have sufficient statistical power to also detect medium sized effects of Cohen's $d = 0.71$ (Power = 0.80, alpha = 0.05, two-tailed independent t-test). All participants gave their consent for participation and were informed about their participation rights. They were paid £2.85 for completing the study. According to the German Research Foundation's guidelines for the Social Sciences and Humanities and the German Association for Psychology's guidelines for ethical acting

in psychological research, the present study did not require ethics approval as it followed standard procedures (e.g., it did not involve clinical populations or deception of participants, and posed no risks to the participants). Yet, we would like to point out that a follow-up study that closely resembles the present one was approved by the Ethics Committee of the Institute of Philosophy and Educational Science of the Ruhr University Bochum (EPE-2023– 005).

2.2 Experimental Procedure

Participants performed the same dot-matching task as in Experiment 1 in Samson et al. [30] but with a robot avatar instead of a human avatar. In each trial, participants first saw a fixation cross (500 ms), followed by a 500 ms blank screen, and then either the word "You" or "Robot" (750 ms), specifying from which perspective the task should be performed for the present trial. Then, after an additional blank of 500 ms, a digit (0–3) was presented (750 ms), followed by a picture of a room, in which a varying number of red dots were presented (0–3) and the robot avatar. The picture was presented till the participant made a response. The participant's task was to indicate whether the digit matches the number of red dots that can be seen from the perspective specified beforehand (either You or Robot). The red dots were presented either in a way that they match or mismatch the digit from the specified perspective. Moreover, the red dots seen by the participant could either be the same as seen for the robot (consistent trials) or different (inconsistent trials) – for an overview of all trial types, see Fig. 1. That is, the robot can see the dots on the wall that it is facing but not the dots behind it on the other wall. The participant, however, can see the dots on *both* walls. Thus, the number of dots that can be seen by the robot and participant can be consistent (e.g., if two red dots are presented on the wall that the robot is facing) or inconsistent (e.g., if one red dot is presented on each wall). Participants could either respond that the dots match the digit from the specified perspective by pressing the 'm' key or mismatch by pressing the 'n' key on their keyboard. They were instructed to respond as accurately and as fast as possible.

Importantly, the dot-matching task allows to measure egocentric as well as altercentric intrusions [30]. Egocentric intrusions are quantified in trials, in which participants are required to take the perspective of the robot. Here, a slowing down of response times for inconsistent vs. consistent trials is attributed to an interference from the participant's own perspective. Conversely, altercentric intrusions are quantified in trials, in which participants are required to take their own perspective. Here, a slowing down of response times for inconsistent vs. consistent trials is attributed to an interference from the robot's perspective.

Prior to performing the actual experiment, each participant performed 8 practice trials to get familiar with the task. In the actual experiment, each participant performed 156 trials, which were composed of 36 consistent matching trials, 36 inconsistent matching trials, 36 consistent mismatching trials, 36 inconsistent mismatching trials, and 12 filler trials, and presented in a random order. For each set of 36 trials, half of the trials required participants to perform the task from their own perspective and half from the robot's perspective. In the filler trials, as in Samson et al. [30], participants saw zero red dots on the walls. Half of the participants saw the robot facing the left wall in all trials, and the other half of participants saw the robot facing the right wall in all trials. The experiment took about 19 min to complete. It was programmed in *PsychoPy* [29]. All analyses were performed using custom *R* scripts.

Depending on the condition and prior to performing the practice trials, participants would be either instructed that the robot is turned on or off. The exact wording for the on-instruction was (for participants seeing the robot facing the right wall; phrasings were analogous for the robot facing the left wall): "Importantly, the ROBOT is turned ON, which is indicated by a green light on the shoulder of the ROBOT. It can see the red dots on the right wall." The exact wording for the off instruction was (again, phrasings were analogous for the robot facing the left wall): "Importantly, the ROBOT is turned OFF, which is indicated by a red light on the shoulder of the ROBOT. Even though it is looking to the right, it cannot see the red dots on the right wall. Still, if the ROBOT's perspective is specified, it is your task to judge from its perspective whether the digit matches the number of red dots that the ROBOT would see on the right wall." Also, depending on the experimental condition, participants would either always see a robot with a human-like head or with an artificial-looking head (for an overview of our between-subjects conditions, see Fig. 2).

3 Results

In line with Samson et al. [30], we only analyzed data of matching trials (yes responses). Given that Samson et al. [30] had a response timeout of 2 s and only analyzed correct responses, we also removed all trials, in which participants took longer to respond than 2 s and made incorrect responses (7%).

With the dependent variable response times (see Figs. 3 and 4, for a descriptive overview), we performed a 2 × 2 × 2 × 2 ANOVA with the within-subjects factors Perspective (Self vs. Robot) and Consistency (Consistent vs. Inconsistent) and the between-subjects factors Status (On vs. Off) and Appearance (Human-like vs. Artificial). We found significant main effects of Consistency ($F(1,124) = 12.81$, $p < 0.001$, $\eta^2 = 0.052$; Consistent $M = 737$ ms, Inconsistent $M = 822$ ms) and Perspective ($F(1,124) = 29.85$, $p < 0.001$, η^2

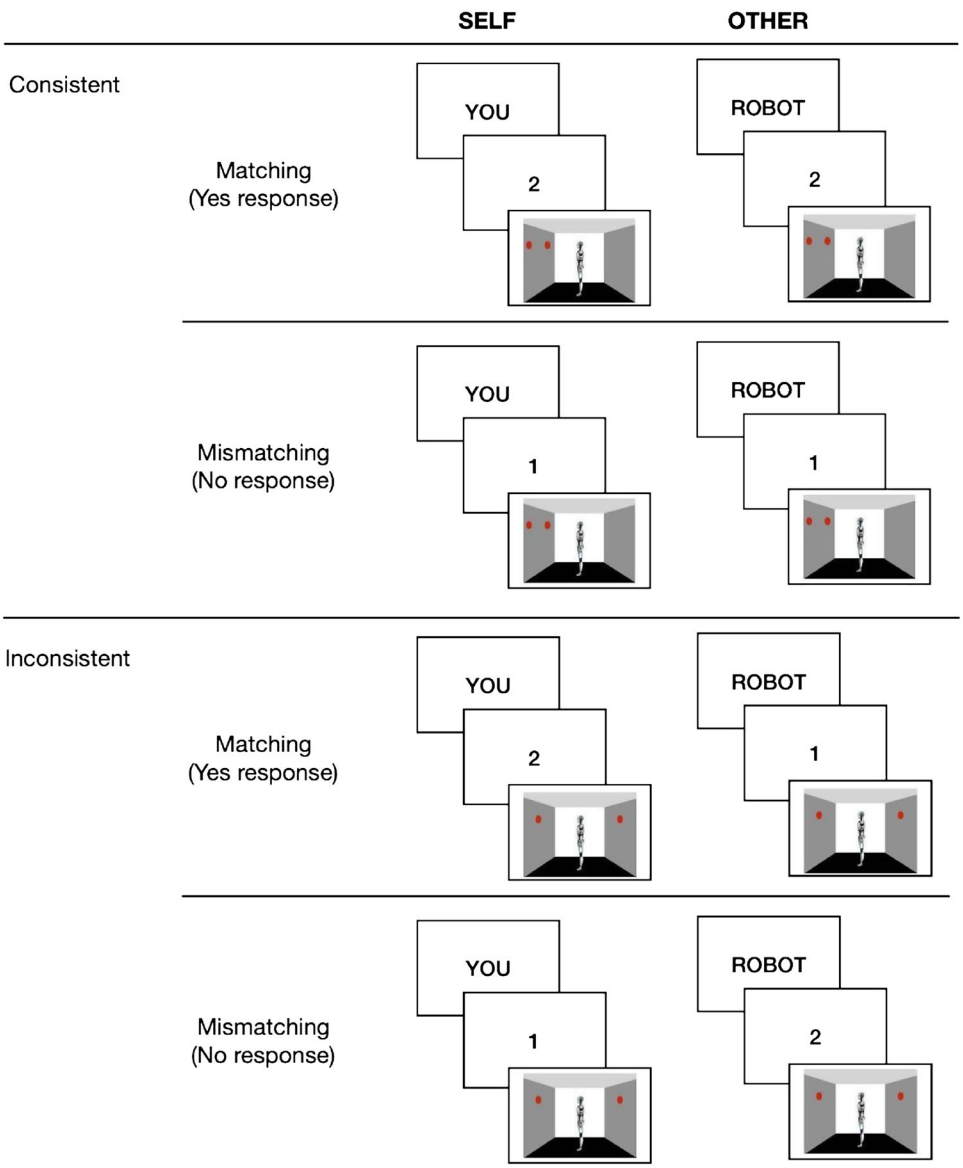

Fig. 1 Example trials for our within-subjects factors: Consistency (Consistent vs. Inconsistent) and Perspective (Self vs. Other)

= 0.009; Robot: $M = 762$ ms, Self: $M = 797$ ms). Moreover, we found a significant interaction effect between these two factors ($F(1,124) = 17.09$, $p < 0.001$, $\eta^2 = 0.005$). The interaction effect between Status and Appearance approached significance ($F(1,124) = 3.72$, $p = 0.056$, $\eta^2 = 0.023$).

All other effects were not significant ($ps > 0.175$, see Table 1 for a list of all effects), suggesting that the Appearance and Status factors had no influence on the Consistency and Perspective factors and thus no influence on altercentric and egocentric intrusion effects. To better quantify the absence of these influences, we additionally computed Bayes Factors for all our effects using a Bayesian ANOVA. The computed Bayes factors are exclusion factors (i.e. the reciprocal of inclusion Bayes' factors [15]) and thus indicate for each effect of the ANOVA how much more likely the null hypothesis is compared to the respective effect. All the Bayes factors assessing the influence of the Appearance and/or the Status factor on the Consistency and/or Perspective factors are greater than 3 (see last column in Table 1), suggesting that the evidence in favor of the null hypothesis is at least substantial [17] for these effects.

The significant interaction effect between Consistency and Perspective mentioned above indicates that the size of the Consistency effect is dependent on the perspective taken by the participant (see Fig. 5, for a descriptive overview). That is, the difference between Consistent and Inconsistent conditions when taking one's own perspective is smaller than when taking the robot's perspective. When testing the Consistency effect only for the data when taking one's own perspective, we find that it is still significant ($t(127) = 6.20$, $p < 0.001$, Cohen's $d = 0.55$, $M = 59.18$ ms, $SE = 9.54$) and thus, importantly, indicates altercentric intrusions. When taking

Table 1 Response times: ANOVA results and exclusion Bayes' factors

Effects	F	df	p	η^2	BF_{excl}
Status	0.324	1	0.570	0.002	6.053
Appearance	1.497	1	0.223	0.009	3.671
Status * appearance	3.723	1	0.056	0.023	3.161
Residuals		124			
Consistency	128.814	1	< 0.001	0.052	8.648×10^{-16}
Consistency * status	1.872	1	0.174	7.544×10^{-4}	3.122
Consistency * appearance	0.953	1	0.331	3.839×10^{-4}	4.236
Consistency * status * appearance	0.880	1	0.350	3.547×10^{-4}	38.648
Residuals		124			
Perspective	29.854	1	< 0.001	0.009	4.583×10^{-7}
Perspective * status	0.050	1	0.823	1.501×10^{-5}	12.187
Perspective * appearance	0.901	1	0.344	2.691×10^{-4}	7.575
Perspective * status * appearance	1.136	1	0.289	3.391×10^{-4}	52.932
Residuals		124			
Consistency * perspective	17.088	1	< .001	0.005	8.113×10^{-4}
Consistency * perspective * status	0.708	1	0.402	1.964×10^{-4}	38.953
Consistency * perspective * appearance	0.277	1	0.600	7.679×10^{-5}	24.835
Consistency * perspective * status * appearance	0.005	1	0.945	1.314×10^{-6}	94,699.783
Residuals		124			

the robot's perspective, the Consistency effect is significant as well and thus indicates egocentric intrusions ($t(127) = 11.22$, $p < 0.001$, Cohen's $d = 0.99$, $M = 110.45$ ms, $SE = 9.85$). When directly comparing these intrusions, we find that the egocentric intrusion effect is significantly larger than the altercentric intrusion effect ($t(127) = 4.17$, $p < 0.001$, Cohen's $d = 0.37$), in line with Samson et al. [30]. This difference in magnitude of the intrusion effects appears to result from faster response times in the consistent conditions when taking the robot's perspective compared to taking one's own perspective. Indeed, the consistent conditions significantly differ between perspectives ($t(127) = 8.30$, $p < 0.001$, Cohen's $d = 0.73$) but do not differ for the inconsistent conditions between perspectives ($t(127) = 0.93$, $p = 0.353$, Cohen's $d = 0.08$).

The interaction effect between Status and Appearance mentioned above, which approached significance, appears to be mainly driven by faster response times when the robot has a human-like appearance and is turned off ($M = 722.78$ ms, $SE = 30.08$) compared to when it is turned on ($M = 793.88$ ms, $SE = 26.87$). This difference in response times is smaller when the Robot's appearance is artificial (ON: $M = 774.63$ ms, $SE = 30.49$ vs. OFF: $M = 816.43$ ms, $SE = 29.83$). However, given that this effect only approached significance, we refrain from any further interpretation and suggest that future studies are needed to test whether this effect is statistically reliable.

When repeating the same $2 \times 2 \times 2 \times 2$ ANOVA with accuracy as the dependent variable (see Figs. 6 and 7, for a descriptive overview), we only find a significant main effect of Consistency ($F(1,124) = 102.57$, $p < 0.001$, $\eta^2 = 0.135$; Consistent $M = 0.98$, Inconsistent $M = 0.92$) but no other significant effects ($ps > 0.085$; see Table 2 for a list of all effects and exclusion Bayes Factors), suggesting our response time results cannot be alternatively explained by a speed-accuracy trade-off.

4 Discussion

4.1 The Mere-Appearance Hypothesis and Robotic Avatars

The current study aims to contribute to the debate regarding VPT when humanoid robotic agents are involved. In particular, we addressed the difference in results by Zhao and Malle [45], who finds VPT2, and Xiao et al [43], who does not find VPT1. To address these diverging findings, we tested whether VPT1 is triggered using (as [43] the paradigm by Samson et al. [30] but, contrary to [43], with humanoid robot avatars, possessing clearly identifiable features indicating the direction of sight. Moreover, we varied the appearance of the robot (artificial head vs human-like head) to test the robustness of the mere-appearance hypothesis proposed by Zhao

Table 2 Accuracy: ANOVA results and exclusion Bayes' factors

Effects	F	df	*p*	η^2	BF_{excl}
Status	0.059	1	0.809	1.585×10^{-4}	40.156
Appearance	1.339	1	0.249	0.004	21.651
Status * appearance	0.991	1	0.321	0.003	74.745
Residuals		124			
Consistency	102.571	1	< .001	0.135	0.000
Consistency * status	0.073	1	0.788	9.588×10^{-5}	41.508
Consistency * appearance	0.179	1	0.673	2.368×10^{-4}	24.949
Consistency * status * appearance	0.001	1	0.969	1.957×10^{-5}	1849.119
Residuals		124			
Perspective	0.059	1	0.809	9.588×10^{-5}	33.009
Perspective * status	0.271	1	0.604	4.402×10^{-4}	141.131
Perspective * appearance	0.098	1	0.755	1.585×10^{-4}	89.239
Perspective * status * appearance	0.531	1	0.468	8.629×10^{-4}	5291.362
Residuals		124			
Consistency * perspective	3.013	1	0.085	0.004	12.025
Consistency * perspective * status	0.367	1	0.546	4.402×10^{-4}	1812.337
Consistency * perspective * appearance	0.080	1	0.778	9.588×10^{-5}	1412.696
Consistency * perspective * status * appearance	2.478	1	0.118	0.003	$4.531 \times 10^{+6}$
Residuals		124			

and Malle [45] and how the robot was presented (switched on vs. off) to assess whether a mental capacity to actually see objects is necessary for VPT1 to occur.

Contrary to Xiao et al. [43], we do find altercentric intrusions, suggesting that VPT1 is triggered also for humanoid robots. This is in line with the VPT2 effect found in the study by Zhao and Malle [45], since human-like appearance seems to be sufficient to trigger VPT1 after all. In addition, presenting the robot as turned "off" did not eliminate altercentric intrusions, suggesting that the human-like appearance of the robot seems to be sufficient and information about its mental capacity to actually see an object is not required, providing further support for the mere-appearance hypothesis. Also, varying the appearance of the robot did not affect altercentric intrusions, suggesting that the mere-appearance hypothesis, which rests on the idea that human-like appearance is enough to trigger VPT, is quite robust when it comes to big alterations to the human-like appearance (i.e., using an artificial camera head vs. human-like head). These findings, however, suggest that more investigation is needed when it comes to the boundary conditions of the mere-appearance hypothesis, if it is true that human-likeness is sufficient to trigger VPT but a human-like head is not necessary, one must wonder about if there are human features that are, on the contrary, necessary for human-likeness to be detected and, consequently, for VPT to occur, and what they are. Future experiments could thus test what are the basic features of human appearance that trigger VPT.

4.2 Stimuli Matter

Why do our findings differ from Xiao et al. [43]? On the one hand, as already noted in the introduction, it might be a matter of the used stimuli. While with our stimuli the direction of sight of the avatar was clearly visible, this was not the case in Xiao et al. [43]. On the other hand, a difference in experimental manipulations could also explain differing results. That is, our study focused on varying appearance features and on information regarding whether the robot is switched on or off. Xiao et al. [43] compared altercentric intrusion effects between human and robotic avatars. One possibility is that presenting both avatars to the same participants made the difference between humans and robots more salient, prompting the participants to see the robots as not sufficiently human, inhibiting VPT. Yet another possibility is, as already acknowledged by Xiao et al. [43], that the findings might be influenced by cultural differences [19], as participants in Xiao et al. [43] were located in China and participants in our sample were located in the US and UK. Future studies could address these points.

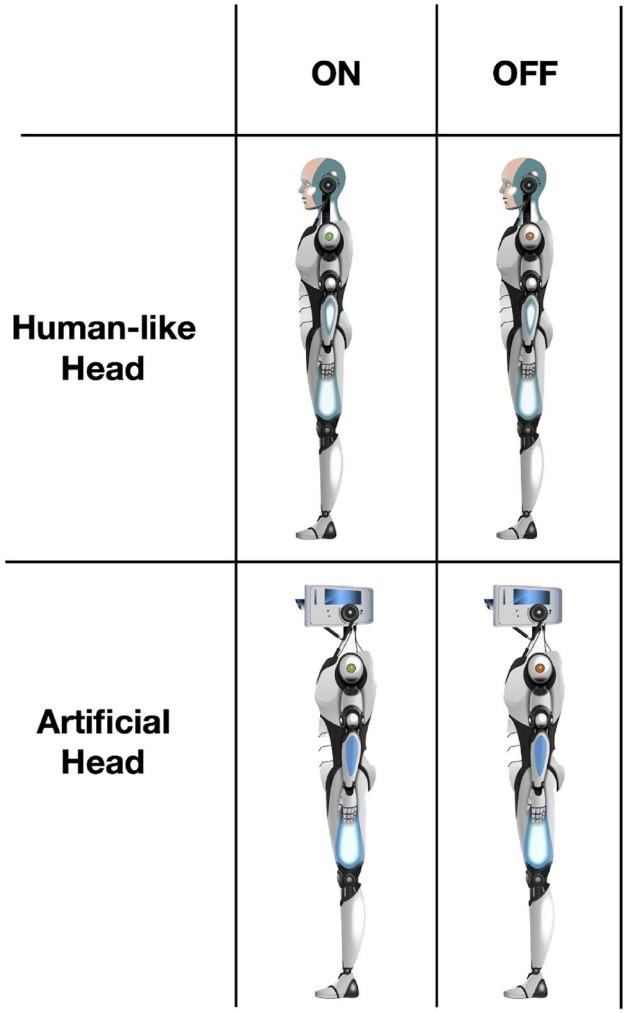

Fig. 2 Overview of our between-subjects factors: Status (On vs. Off) and Appearance (Human-like vs. Artificial)

4.3 The Dot-Matching Task and VPT

It should also be acknowledged that a current debate is concerned with the validity of the dot-matching task as a means of testing VPT and its spontaneity and/or automaticity in the first place. In particular, the debate has focused on whether the dot-matching task in its original format and its variations test VPT or re-orienting of attention due to directional cues (e.g. [11, 20, 44]). Recently, O'Grady et al. [28] reconciled findings in this debate by showing how the differences in earlier findings are likely due to confounding factors in experimental designs. In particular, they showed that part of the discrepancy in the current data, where some experiments support the idea that directional orienting might be what explains intrusions, and others suggest that VPT is

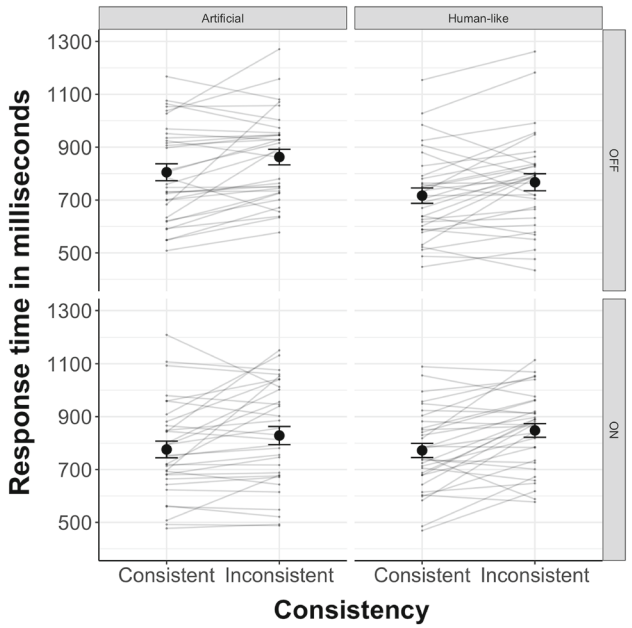

Fig. 3 Descriptive overview for response times for trials, in which participants performed the dot-matching task from their own perspective. The difference between consistent and inconsistent conditions indicate the extent to which the robot's perspective interfered with response times – *altercentric* intrusion effects. Light gray lines indicate individual participants. Error bars are standard error of the mean

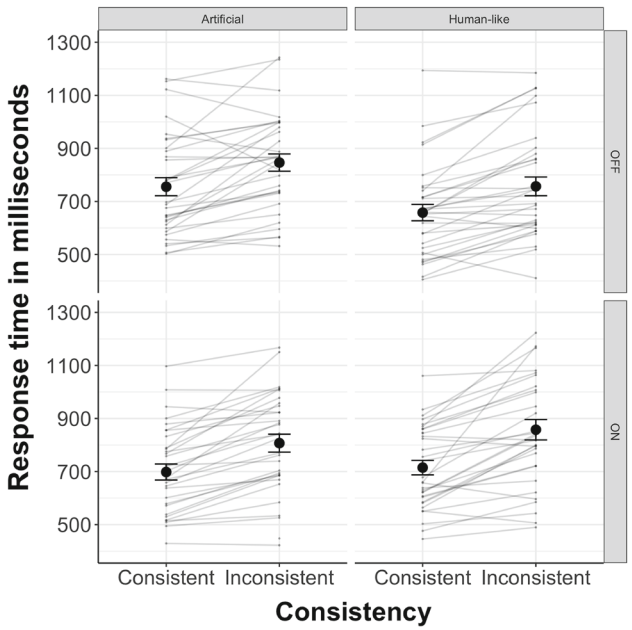

Fig. 4 Descriptive overview for response times for trials, in which participants performed the dot-matching task from the robot's perspective. The difference between consistent and inconsistent conditions indicate the extent to which the participant's own perspective interfered with response times – *egocentric* intrusion effects. Light gray lines indicate individual participants. Error bars are standard error of the mean

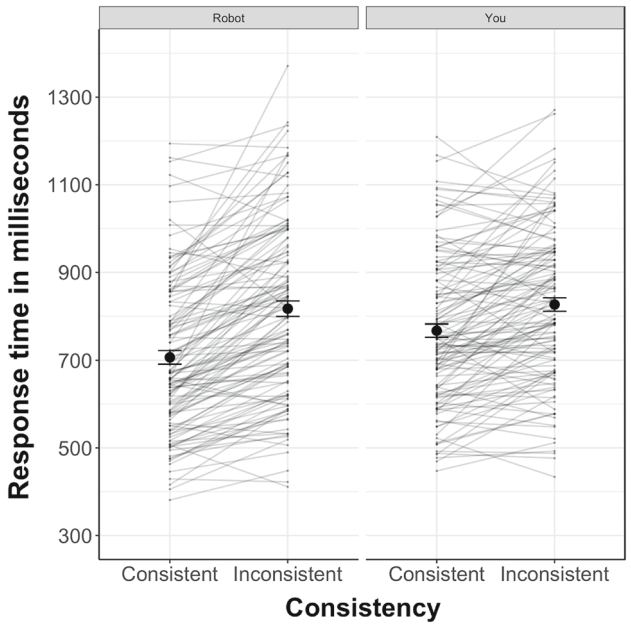

Fig. 5 Descriptive overview for response times for trials, in which participants performed the dot-matching task from the robot's perspective (left panel—egocentric intrusions) and their own perspective (right panel—altercentric intrusions). Light gray lines indicate individual participants. Error bars are standard error of the mean

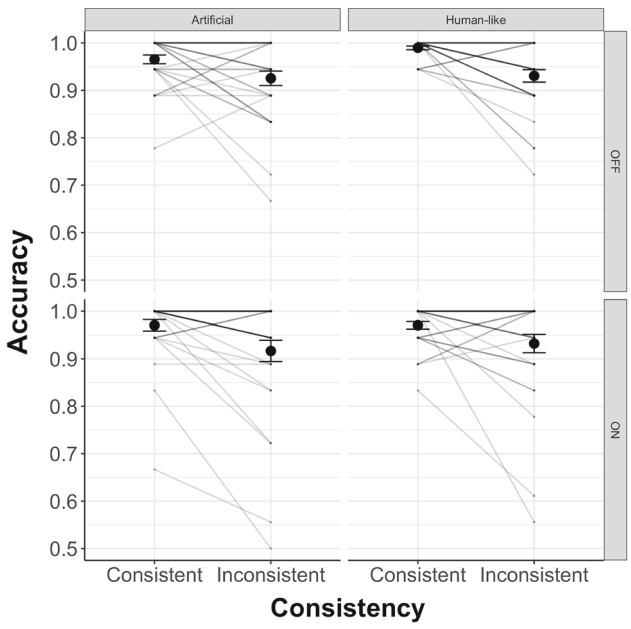

Fig. 6 Descriptive overview for accuracies for trials, in which participants performed the dot matching task from their own perspective. Light gray lines indicate individual participants. Error bars are standard error of the mean

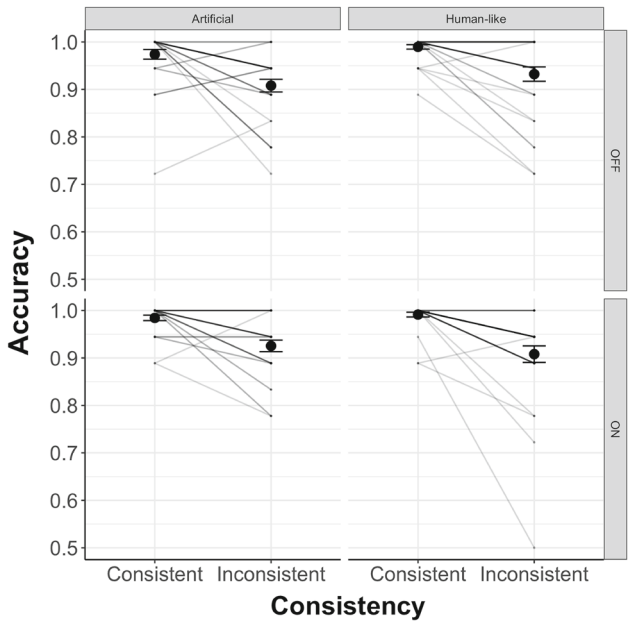

Fig. 7 Descriptive overview for accuracies for trials, in which participants performed the dot matching task from the robot's perspective. Light gray lines indicate individual participants. Error bars are standard error of the mean

involved [3, 9–11, 20, 31, 44] is due to differences in the explicitness of the VPT task. Explicit tasks, where the participant is directly instructed to adopt her own or the avatar's point of view, do indeed show evidence of altercentric intrusions in VPT. Conversely, results in implicit tasks suggest that, without instructions, directional orienting likely drives the found effects. Their experimental findings also support this conclusion. The difference between implicit and explicit task findings are due to the fact, O'Grady and colleagues argue, that VPT does indeed occur rapidly and involuntarily, but only in tasks where VPT is explicitly relevant (and thus, is "spontaneous" rather than "automatic"). As we use here an explicit task, we believe our experiment fits well within this interpretation of their data, and thus suggest that visual perspective taking (rather than directional orienting) is involved when robotic avatars are used.

However, we also want to point out that studies exploring VPT towards robots and comparing it with VPT towards humans can potentially help us narrow down which factors do influence spontaneous VPT or trigger it in the first place. For example, future studies could test the boundaries of the mere-appearance hypothesis further by identifying robot avatars that lack human-likeness (and thus show no altercentric intrusions) but still clearly possess directional cues (e.g., an artificial-looking head looking in a particular

direction). Such avatars would help to identify to what extent directional cues contribute towards VPT effects.

4.4 Study Limitations

With regard to limitations of our study, we want to note that it could be the case that our manipulation for the on and off conditions may have been too subtle in terms of visual cues. That is, the green and red dots at the shoulder of the avatar may not have been sufficient to remind the participants that the robot was turned on or off, respectively, despite our prior instructions. While one could have used more salient visual indicators of the robot's status, the more salient indicators could have diverted attention away from the robot itself. Also, more salient visual indicators could have been confounded with the directional cues of the robot. For instance, a ray of light from the eyes of the robot that is switched on would also constitute an additional directional cue and consequently interfere with the purpose of the manipulation, that is concerned with mental capacity. Conversely, closing the eyes of the robot that is switched off would reduce the directional cues of the robot. Further studies, however, could address these concerns and present a manipulation of mental capacity that is not too salient and does not interfere with directional cues. Possibly, future studies could approach this issue by showing videos to participants of the robot that is switched on, in which it briefly interacts with the participant (e.g., by greeting it), *prior* to performing the dot-matching task.

4.5 Conclusions and Outlook

With robots increasingly entering our social lives, investigating the ways we understand them as agents is of fundamental importance. Whether or not we spontaneously take the perspective of a robot has the strong potential to influence our coordination with them. This is essential if we aim to develop smooth interactions and joint actions [41] as well as successful communication with robots. VPT is especially relevant when it comes to reference fixing in communication [24, 26]. That is, when communicating we often rely on what we think is shared information on what others know and see and the ability to take the perspective of others is critical to infer this information [1, 16, 25]. The mere-appearance hypothesis formulated by Zhao and Malle [45] provides us with important guidance regarding what possible features of robots might be critical when it comes to thinking about robots involved in our social life and easier to coordinate and communicate with. Our study provides further support for this hypothesis and also shows that it is robust with regard to relatively big alterations to human appearance. Future studies could thus further explore this direction of research by further testing the boundary conditions of the mere-appearance hypothesis. Understanding what factors influence perspective taking, including to what extent human-likeness does, has the potential of advancing our understanding of the underlying mechanisms (i.e., to what extent VPT and attentional orienting contribute towards intrusion effects) as well as informing our design and expectations concerning social robots.

Funding Open Access funding enabled and organized by Projekt DEAL. BW and LB were funded from the ministry of culture and science of Northrhine Westphalia (cooperative research project "INTERACT!"). The sole responsibility for the content of this paper lies with the authors.

Data availability All data is available in the following repository: https://osf.io/x76fc/?view_only=095423ecab694936a1a3ba4c6082c701.

Code availability Experimental code and analysis scripts are freely available on request.

Declarations

Conflict of interests The authors have no competing interests to declare.

References

1. Berio L, Vosgerau G (2020) Enriching the cognitive account of common ground: kinds of shared information and cognitive processes. Grazer Philosophischen Studien 97(3):495–527. https://doi.org/10.1163/18756735-000105
2. Clark HH, Fischer K (2022) Social robots as depictions of social agents. Behav Brain Sci. https://doi.org/10.1017/S0140525X22000668
3. Conway JR, Lee D, Ojaghi M, Catmur C, Bird G (2017) Submentalizing or mentalizing in a level 1 perspective-tak- ing task: a cloak and goggles test. J Exp Psychol Hum Percept Perform 43(3):454–465. https://doi.org/10.1037/xhp0000319
4. Cross ES, Hortensius R, Wykowska A (2019) From social brains to social robots: applying neurocognitive insights to human–robot interaction. Philos Trans R Soc B 374(1771):20180024
5. de Graaf MMA, Malle BF (2019) People's explanations of robot behavior subtly reveal mental state inferences. In: 2019 14th ACM/IEEE international conference on human-robot interaction (HRI), Daegu, Korea (South), 2019, pp 239–248. https://doi.org/10.1109/HRI.2019.8673308.
6. Edwards K, Low J (2018) Level 2 perspective-taking distinguishes automatic and non-automatic belief tracking. Cognition 193:104017. https://doi.org/10.1016/j.cognition.2019.104017

7. Epley N, Waytz A, Cacioppo JT (2007) On seeing human: a three-factor theory of anthropomorphism. Psychol Rev 114(4):864–886. https://doi.org/10.1037/0033-295X.114.4.864
8. Flavell JH, Everett BA, Croft K, Flavell ER (1981) Young children's knowledge about visual perception: further evidence for the level 1–level 2 distinction. Dev Psychol 17(1):99–103
9. Freundlieb M, Kovács AM, Sebanz N (2016) When do humans spontaneously adopt another's visuospatial perspective? J Exp Psychol Hum Percept Perform 42(3):401–412. https://doi.org/10.1037/xhp0000153
10. Freundlieb M, Kovács AM, Sebanz N (2018) Reading your mind while you are reading—evidence for spontaneous visuospatial perspective taking during a semantic categorization task. Psychol Sci 29(4):614–622. https://doi.org/10.1177/0956797617740973
11. Gardner MR, Hull Z, Taylor D, Edmonds CJ (2018) "Spontaneous" visual perspective-taking mediated by attention orienting that is voluntary and not reflexive. Q J Exp Psychol 71(4):1020–1029. https://doi.org/10.1080/17470218.2017.1307868
12. Gray K, Wegner DM (2012) Feeling robots and human zombies: mind perception and the uncanny valley. Cognition 125(1):125–130. https://doi.org/10.1016/j.cognition.2012.06.007
13. Guttman N, Kalish HI (1956) Discriminability and stimulus generalization. J Exp Psychol 51(1):79–88
14. Hamilton AFDC, Brindley R, Frith U (2009) Visual perspective taking impairment in children with autistic spectrum disorder. Cognition 113(1):37–44
15. Hinne M, Gronau QF, van den Bergh D, Wagenmakers EJ (2020) A conceptual introduction to Bayesian model averaging. Adv Methods Pract Psychol Sci 3(2):200–215
16. Horton WS, Gerrig RJ (2005) Conversational common ground and memory processes in language production. Discourse Process 40(1):1–35
17. Jeffreys H (1961) Theory of probability, 3rd edn. Oxford University Press, Oxford, UK
18. Kessler, K., & Rutherford, H. (2010). The two forms of visuospatial perspective taking are differently embodied and subserve different spatial prepositions. Front Psychol 1:213
19. Kessler K, Cao L, O'Shea KJ, Wang H (2014) A cross-culture, cross-gender comparison of perspective taking mechanisms. Proc R Soc B 281(1785):20140388
20. Langton S (2018) I don't see it your way: the dot perspective task does not gauge spontaneous perspective taking. Vision 2(1):6. https://doi.org/10.3390/vision2010006
21. Martin AK, Perceval G, Davies I, Su P, Huang J, Meinzer M (2019) Visual perspective taking in young and older adults. J Exp Psychol Gen 148(11):2006–2026. https://doi.org/10.1037/xge0000584
22. Mathur MB, Reichling DB (2016) Navigating a social world with robot partners: a quantitative cartography of the Uncanny Valley. Cognition 146:22–32. https://doi.org/10.1016/j.cognition.2015.09.008
23. Meltzoff AN, Brooks R, Shon AP, Rao RP (2010) "Social" robots are psychological agents for infants: a test of gaze following. Neural Netw 23(8–9):966–972. https://doi.org/10.1016/j.neunet.2010.09.005
24. Mozuraitis M, Stevenson S, Heller D (2018) Modeling reference production as the probabilistic combination of multiple perspectives. Cogn Sci 42:974–1008. https://doi.org/10.1111/cogs.12582
25. Nadig AS, Sedivy JC (2002) Evidence of perspective-taking constraints in children's on-line reference resolution. Psychol Sci 13(4):329–336. https://doi.org/10.1111/j.0956-7976.2002.00460.x
26. Newen A (1998) Reference and reference determination: the interpretational theory. Lingua Et Style 33:515–529
27. Nielsen KM, Slade L, Levy JP, Holmes A (2015) Inclined to see it your way: Do altercentric intrusion effects in visual perspective taking reflect an intrinsically social process? Q J Exp Psychol (Hove) 68(10):1931–1951. https://doi.org/10.1080/17470218.2015.1023206
28. O'Grady C, Scott-Philipps T, Lavelle S, Smith K (2020) Perspective taking is spontaneous but not automatic. Q J Exp Psychol 73(10):1605–1628
29. Peirce J, Hirst R, MacAskill M (2022) Building experiments in PsychoPy. Sage
30. Samson D, Apperly IA, Braithwaite JJ, Andrews BJ, Bodley Scott SE (2010) Seeing it their way: evidence for rapid and involuntary computation of what other people see. J Exp Psychol Hum Percept Perform 36(5):1255–1266. https://doi.org/10.1037/a0018729
31. Santiesteban I, Catmur C, Hopkins SC, Bird G, Heyes C (2014) Avatars and arrows: Implicit mentalizing or domaingeneral processing? J Exp Psychol Hum Percept Perform 40(3):929–937. https://doi.org/10.1037/a0035175
32. Shepard RN (1987) Towards a universal theory of generalization for psychological science. Science 237(4820):1317–1323
33. Sheridan TB (2016) Human–robot interaction: status and challenges. Hum Factors 58(4):525–532. https://doi.org/10.1177/0018720816644364
34. Singh SJ, Kapoor DS, Sohi BS (2021) All about human-robot interaction. In: Mittal M, Shah RR, Roy S (eds), Cognitive data science in sustainable computing, cognitive computing for human–robot interaction. Academic Press, pp 199–229. https://doi.org/10.1016/B978-0-323-85769-7.00010-0
35. Surtees A, Apperly I, Samson D (2013) Similarities and differences in visual and spatial perspective-taking processes. Cognition 129(2):426–438. https://doi.org/10.1016/j.cognition.2013.06.008
36. Surtees A, Samson D, Apperly I (2016) Unintentional perspective-taking calculates whether something is seen, but not how it is seen. Cognition 148:97–105
37. Todd AR, Cameron CD, Simpson AJ (2021) The goal-dependence of level-1 and level-2 visual perspective calculation. J Exp Psychol Learn Mem Cogn 47(6):948–967. https://doi.org/10.1037/xlm0000973
38. Tomasello M (1999) The cultural origins of human cognition. Harvard University Press
39. Tomasello M (2018) Becoming human: a theory of ontogeny. Harvard University Press
40. Urgen BA, Plank M, Ishiguro H, Poizner H, Saygin AP (2013) EEG theta and Mu oscillations during perception of human and robot actions. Front Neurorobot 13(7):19. https://doi.org/10.3389/fnbot.2013.00019
41. Vesper C, Abramova E, Bütepage J, Ciardo F, Crossey B, Effenberg A, Hristova D, Karlinsky A, McEllin L, Nijssen S, Schmitz L, Wahn B (2017) Joint action: mental representations, shared information and general mechanisms for coordinating with others. Front Psychol 7:2039. https://doi.org/10.3389/fpsyg.2016.02039
42. Waytz A, Norton MI (2014) Botsourcing and outsourcing: robot, British, Chinese, and German workers are for thinking—not feeling—jobs. Emotion 14(2):434–444. https://doi.org/10.1037/a0036054
43. Xiao C, Fan Y, Zhang J, Zhou R (2022) People do not automatically take the level-1 visual perspective of humanoid robot avatars. Int J Soc Robot 14(1):165–176
44. Zhao X, Cusimano C, Malle BF (2015) In search of triggering conditions for spontaneous visual perspective taking. In Noelle DC, Dale R, Warlaumont AS, Yoshimi J, Matlock T, Jennings CD, Maglio PP (Eds), Proceedings of the 37th annual meeting of the cognitive science society (pp 2811–2816). Cognitive Science Society.
45. Zhao X, Malle BF (2022) Spontaneous perspective taking toward robots: the unique impact of humanlike appearance. Cognition 224:105076

Basil Wahn is a Postdoctoral Researcher at Ruhr University Bochum in Germany within the "INTERACT!" project. In his research, he investigates the mechanisms that enable smooth and efficient interactions between humans and between humans and artificial agents. Specifically, he is interested in the factors that influence the willingness of humans to collaborate with, trust, and take the perspective of artificial agents.

Leda Berio is a Postdoctoral Researcher at Ruhr University Bochum in Germany within the "INTERACT!" project and also a collaborator within the "The Communicative Mind" project at the University of Warwick. In her research, she investigates the impact of language and culture on our folk psychology, on our perception of gender roles as well as racialised groups, and on interaction with artificial agents.

Matthias Weiss is Professor of Management and Head of the ZEPPELIN Chair of Innovation Management and Transformation at Zeppelin University in Friedrichshafen (Germany). Before joining Zeppelin University, he served on the faculties of Radboud University Nijmegen, Ruhr-University Bochum, and Ludwig-Maximilians-University Munich. Moreover, he was a visiting researcher at Bocconi University (Milan, Italy) and at the University of Lugano (Switzerland). His main research interests are located at the intersection between innovation management and psychology, with a particular focus on the consequences of digital technologies for creative (team)work.

Albert Newen is currently full professor of philosophy at the Ruhr University Bochum, Germany. His research fields are philosophy of mind and cognition with a focus on theories of self and agency, social understanding, emotion and perception. The methodological perspective includes a focus on situated cognition as well as comparative perspective for humans, animals and robots.