Cognition 259 (2025) 106101

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Good to see you R2-D2: Inducing spontaneous perspective-taking towards non-human agents through human-like gaze and reach

Xucong Hu, Haokui Xu, Hui Chen *, Mowei Shen *, Jifan Zhou *

Department of Psychology and Behavioral Sciences, Zhejiang University, China

ARTICLE INFO

Keywords:
Spontaneous perspective taking
Human-agent interaction
Stimulus-response compatibility

ABSTRACT

Observing the world from another's perspective is a fundamental social cognitive ability essential for human cooperation. With the increasing prevalence of intelligent systems in our society, highly intelligent social robots such as R2-D2 in Star Wars is becoming a reality, thus it is compelling to explore how this capability can extend from humans to non-human agents. Although previous research indicates that a human-like appearance might facilitate this extension, our study contends that human-like actions are more critical. We conducted four experiments involving agents that did not resemble humans but could perform two human-like actions: reach and gaze, which exhibited the perceptual and behavioral abilities that were essential for social interaction. The experiments found that agents prompted spontaneous perspective-taking among participants when they displayed both actions. Importantly, perspective-taking was maintained only when gaze preceded reach, underscoring the causal relationship that behavior should be interpreted as the consequence of perception. These results highlight the importance of human-like actions rather than mere appearance in fostering spontaneous perspective-taking towards non-human agents, providing insights for improving human-agent interaction.

In *Star Wars: Return of the Jedi*, Luke Skywalker found himself entrapped and on the brink of ejection from a spacecraft. Amidst his despair, R2-D2 emerged atop an adjacent ship, signaling persistently with red lights. After an initial moment of bewilderment, Luke suddenly discerned that from R2's perspective, it had an unobstructed view of his location, and could easily toss the lightsaber to him – that is why it was flashing lights to capture his attention. Upon grasping its intention, Luke successfully retrieved it and escaped (Marquand, 1997). This ability, demonstrated in the scene, is known as *perspective-taking*, which refers to the ability to adopt others' perspectives and infer what they are seeing, feeling and thinking (Flavell, 1977; Samson et al., 2010). As a pivotal social cognition ability, perspective-taking is considered the foundation for human joint action in pursuit of collective goals, as shown in the above example of "transferring the lightsaber": where Luke and R2 interpret each other's perspective to assess the situation (e.g., R2 was able to provide support), facilitating collaboration even without verbal communication (Tomasello, 2008). Notably, as intelligent systems (e.g. robots) increasingly appear in films, video games, and realistic interactive scenarios, people can spontaneously adopt the perspectives of those non-human agents, resembling the interaction with R2-D2 (Freina et al., 2017; Marquand, 1997; Salm-Hoogstraeten and Müsseler, 2021; Xiao et al., 2021). This phenomenon prompts the inquiry into how the social cognitive ability - spontaneous perspective-taking extends from human to non-human agents.

Research on this cognitive process traditionally examines how people take the perspective of "other humans", delineating two primary levels: Level-1, which pertains to the ability to assess *whether* another individual can see a specific object, and Level-2, which involves evaluating *how* an object might appear to that individual (Flavell et al., 1981). Previous research has identified perspective-taking as a *spontaneous* process, indicating that people readily bring themselves into other's perspective in scenarios where other people are merely present, even though not explicitly instructed to do so (Samson et al., 2010; Surtees et al., 2016; Tversky and Hard, 2009; Zhao et al., 2015a). In Samson et al.'s (2010) dot counting task for measuring Level-1 perspective taking, the participants were presented with a scene with another person standing in a room, facing one side of the wall, and different numbers of red dots appearing on each wall (Fig. 1). When asked to judge the number of red dots from their own perspective, their performance decreased if the person saw a different number of dots than they did. The

* Corresponding authors at: Department of Psychology and Behavioral Sciences, Zhejiang University, Zijingang Campus, 866 Yuhangtang Road, Hangzhou 310058, China.

E-mail addresses: chenhui@zju.edu.cn (H. Chen), mwshen@zju.edu.cn (M. Shen), jifanzhou@zju.edu.cn (J. Zhou).

https://doi.org/10.1016/j.cognition.2025.106101
Received 18 July 2024; Received in revised form 26 January 2025; Accepted 26 February 2025
Available online 6 March 2025

results suggest that the participants spontaneously adopted the other person's perspective, thereby being influenced, due to the conflicted contents (dot number) arising from different perspectives. Additionally, Zhao et al. (2015b) provided evidence for Level-2 spontaneous perspective-taking through an alternative paradigm known as the ambiguous number task. Specifically, participants could see a figure "9″ on a table, which appeared as "6″ when seen from the opposite side. This ambiguity of the figure could result in two distinct responses. Notably, participants' responses were slower and they were more inclined to report "6″, when another person was merely presented on the opposite side, in comparison to when the side was unoccupied. Though the other person was irrelevant to the task, participants still incorporated their perspective.

Recent studies have employed the spatial stimulus-response (SR) compatibility paradigms to assess spontaneous perspective-taking in individuals (Böffel and Müsseler, 2019a; Freundlieb et al., 2016; Garofalo et al., 2022; von Salm-Hoogstraeten et al., 2020). In classical spatial stimulus-response (SR) compatibility tasks, participants typically respond to stimuli based on irrelevant attributes such as color or shape. However, their performance enhances when the stimulus location aligns spatially with the intended response (e.g., *left* key for *left* stimulus), a phenomenon known as the SR compatibility effect (Simon, 1969). Interestingly, the effect persists in joint actions (Böffel and Müsseler, 2018; Böffel and Müsseler, 2019b; Freundlieb et al., 2016; Freundlieb et al., 2017; von Salm-Hoogstraeten et al., 2020). For instance, in the avatar-Simon task conducted by Böffel and Müsseler (2019a), participants responded to stimuli based on color (e.g. pressing the *left* key when a *light blue* stimulus appeared) (Fig. 2). Their performance improved when responses were spatially compatible with the stimulus' position relative to an avatar (e.g., pressing the *left* key for a stimulus to the avatar's *left*) compared to incompatible pairings (e.g., pressing the *left* key for a stimulus to the avatar's *right*). This suggests that participants adopt the perspective of the avatar, linking stimulus location from the avatar's perspective with their responses.

The above research demonstrates how people adopt others' perspectives and adjust their responses accordingly, whereas some studies suggest that this effect also extends to *non-human* agents (Dolk et al., 2013; Garofalo et al., 2022; Pick et al., 2014; Salm-Hoogstraeten and Müsseler, 2021). For example, Salm-Hoogstraeten and Müsseler (2021) demonstrated that this effect persisted even when cooperating with a Lego robot. They posit that any salient agent within a shared spatial dimension (e.g. both vary in a horizontal plane) can be adopted as a reference point, thereby influencing people's response. Similarly, Zhao and Malle (2022) found that people spontaneously adopted perspectives of humanoid robots, such as NAO and Baxter. This was attributed to the theory of generalization, which suggests that responses to familiar stimuli (e.g. a human) extend to similar novel stimuli (e.g. a non-human agent) (Guttman and Kalish, 1956; Shepard, 1987). They suggested that the more human-like an agent appears, the more likely people are to adopt its perspective.

Despite substantial evidence supporting spontaneous perspective-taking towards non-human agents, some results remain contentious. For instance, Zhao et al. (2016) reported that even salient agents without eyes did not trigger spontaneous perspective-taking. Similarly, Furlanetto et al. (2013) and Xiao et al. (2022) observed no perspective-taking towards highly human-like agents. Conversely, other studies have observed this phenomenon in agents that resemble humans, suggesting that appearance may not be a critical factor (Carlson et al., 2014; Wahn et al., 2023; Wahn and Berio, 2023; Zhao and Malle, 2022). These inconsistencies prompt a reexamination of the initial question: what are the key factors that elicit spontaneous perspective-taking towards non-human agents?

Through a careful review of this line of research, two key characteristics among non-human agents that effectively prompt perspective-taking emerged. On the one hand, these agents typically possess a distinct visual system, indicating their ability to *perceive* the environment, which includes eye-like structures (Garofalo et al., 2022; Ye et al., 2023), camera-like heads (Wahn et al., 2023; Wahn and Berio, 2023), and even geometrically shaped visual organs (Clements-Stephens et al., 2013). In contrast, agents lacking visible perceptual features often fail to evoke perspective-taking (Furlanetto et al., 2013; Xiao et al., 2022; Zhao et al., 2016). For instance, in the dot-counting task by Xiao et al. (2022), the agent resembled a human, but its face and head were uniformly colored, making it challenging to determine its line of sight. Consequently, participants did not take the agent's perspective in this scenario.

On the other hand, these agents are often endowed with the ability to *behave*, implying that they can interact with the environment and respond to others (Abrini et al., 2023; Müller et al., 2015; Salm-Hoogstraeten and Müsseler, 2021; Ye et al., 2023; Zhao et al., 2016; Zhao and Malle, 2022). For instance, in the experiments conducted by Salm-Hoogstraeten and Müsseler (2021), participants spontaneously adopted the agent's perspective when it reached towards the target after their correct response. Conversely, when the agent did not respond accordingly, participants would not adopt its perspective.

Understanding why perceptual and behavioral capabilities are critical involves recognizing their function in joint actions. Spontaneous perspective-taking aids in interpreting information from others' perspectives, preparing for potential collaboration (Phillips, 2021). The intelligent system that can effectively cooperate with us requires at least two fundamental components: a perceptual module to gather information (input) about the environment and other agents; and a behavioral module to plan and execute actions (output) based on perceived information (Emery, 2000; Gray et al., 2007; Gray and Wegner, 2012). Furthermore, agents behave according to their perception. Therefore,

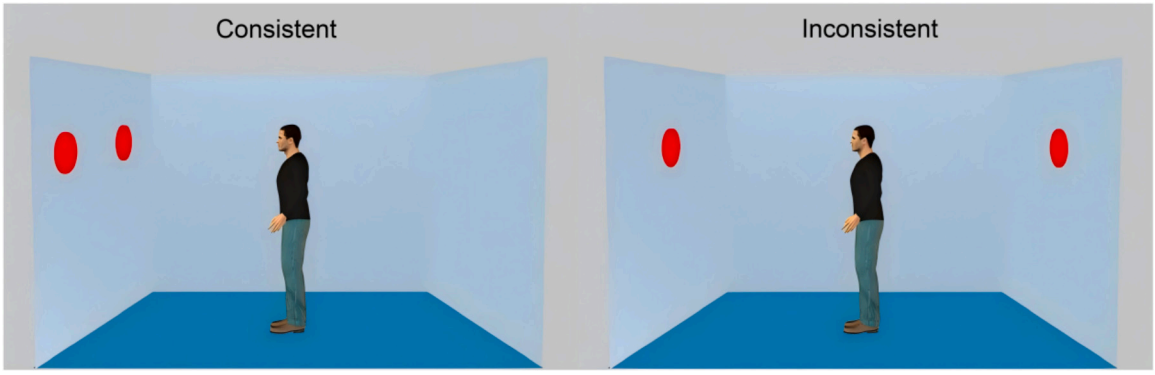

Fig. 1. The schematic illustration of dot-counting task.
Note. The figure illustrates the consistent condition (left) and inconsistent condition (right) in the dot perspective task (Samson et al., 2010). In the consistent condition, the number of dots visible from the avatar's perspective matches the participant's perspective, while in the inconsistent condition, they do not match.

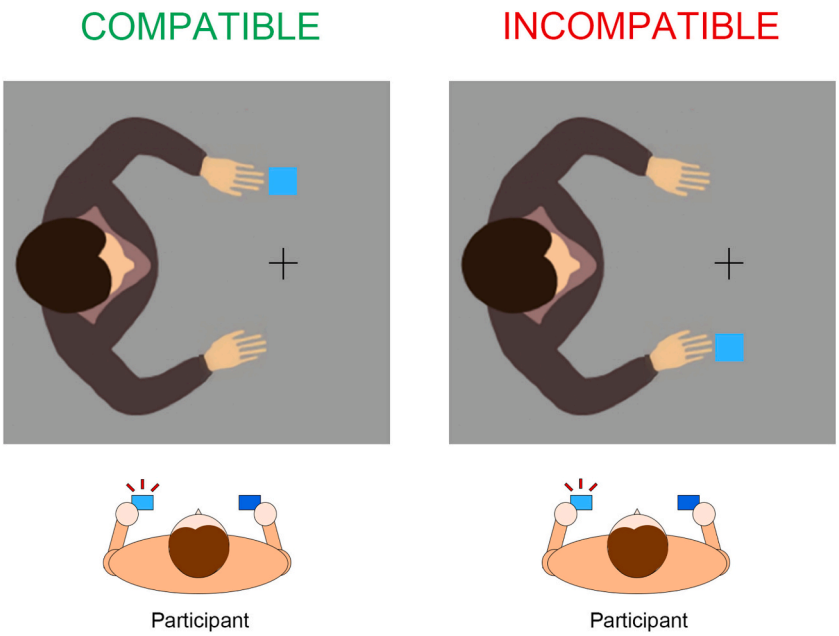

Fig. 2. The schematic illustration of avatar-Simon task.
Note. The figure demonstrates the compatible condition (left) and incompatible condition (right) in the avatar-Simon task (Böffel and Müsseler, 2018). "Compatible" refers to the target location relative to the avatar matching the key location relative to the participants, while "Incompatible" refers to them not matching.

the *causal relationship* between these modules is essential, as perceptual input must precede and inform behavioral output, ensuring appropriate responses and interactions.

In light of the preceding analysis, we posit that perceptual and behavioral abilities of non-human agents are pivotal for eliciting people's spontaneous perspective-taking. Moreover, only when the agent's perception is the antecedent to behavior (i.e., causes precede effects) will spontaneous perspective-taking persist.

This study employed the avatar-Simon task designed by Böffel and Müsseler (2019b), to examine the role of an agent's perceptual and behavioral abilities in eliciting people's spontaneous perspective-taking. Specifically, we utilized two actions – reach and gaze, respectively as the indicators of perceptual and behavioral abilities, which previous research has identified as crucial goal-directed actions also with significant implications for these abilities (Emery, 2000; Tversky and Hard, 2009; Woodward, 1998; Zhao et al., 2015a). Moreover, all agents in this study were intentionally designed to appear distinctly non-human to discount the possibility that a human-like appearance could trigger spontaneous perspective-taking. This design aimed to demonstrate that human-like actions[1] are also pivotal for inducing perspective-taking rather than mere appearance.

Experiment 1 introduced three unique avatars, none resembling humans, each performing one or more actions: reach, gaze, or both. This setup was used to identify under what conditions spontaneous perspective-taking would manifest. Experiment 2 replicated these findings using avatars with a more mechanical appearance to further mitigate any influence of human-like appearance. Experiment 3 then examined how the sequence of gaze and reach (Gaze-Reach and Reach-Gaze) influenced spontaneous perspective-taking, determining the conditions under which perspective-taking persisted. Experiment 4 sought to generalize the findings from Experiments 1 and 2 by applying them to the traditional dot-counting task paradigm established by Samson et al. (2010). This experiment also aimed to rule out alternative explanations, such as the effects of immersion or action control.

1. Experiment 1

Experiment 1 was modified from Böffel and Müsseler (2019a) avatar-Simon task by creating three conditions in which the avatar performed perceptual actions: Reach, Gaze, Gaze&Reach. The objective was to determine whether a non-human avatar capable of both reaching and gazing could effectively elicit spontaneous perspective-taking among people.

1.1. Method

1.1.1. Transparency and openness

For all experiments, we describe how the sample size was determined, detail any data exclusions, and outline all manipulations and measures used in the study. The data sets generated and analyzed during this study are available on the Open Science Framework (OSF): https://osf.io/3dypm/. The design and analysis of this study were not preregistered.

1.1.2. Participants

Thirty-six participants (26 females; age: $M = 20.44$, $SD = 1.18$) were recruited and compensated with 30 CNY or course credit for their participation. All participants in this experiment had normal or corrected-to-normal vision, and signed informed consent to the terms of data collection, usage, and storage.

The required sample size was calculated using the G*Power 3.1 software (Faul et al., 2007). Drawing on the effect sizes observed in previous studies (Böffel and Müsseler, 2019b), we conducted power analyses with an effect size of 0.25, an alpha level of 0.05, a power of 0.80, and a 3×2 within-subjects design. The analyses indicated that 19 participants were required. To ensure thorough counterbalancing across multiple variables, a final sample size of 36 was chosen for Experiments 1 and 2. The study was approved by Institutional Review Board at the Department of Psychology pf the authors' university.

[1] Human-like actions are those that are intention-driven and consistent with human movement patterns. These goal-oriented behaviors, such as gaze and reach, serve to convey perceptual and behavioral intentions, distinguishing them from simple, stimulus-driven responses.

1.1.3. Apparatus and stimuli

The stimuli were generated using Psychopy (version 2023.2.3) (Peirce, 2007), and displayed on a 16-in. monitor with a resolution of 1920 × 1080 pixels. Participants were seated approximately 60 cm away from the monitor and responded with their left and right index fingers on "q" and "p" buttons of the keyboard, each located 8 cm from the participant's midline.

In accordance with the experimental design by Böffel and Müsseler (2019a), we utilized a similar color scheme, but adjusted the size of the target to accommodate our screen resolution. The targets consisted of dark blue squares (RGB 36115254) and light blue squares (RGB 98193254), with the length of 77 pixels in each side (1.34°). These targets were positioned 147 pixels (2.55°) below or above a central fixation cross, set against a gray background (RGB 155155155) measuring 1677 × 1258 pixels (29.26° × 21.94°).

This experiment adopted several simple shapes to form an avatar that does not look like human at all. Meanwhile, the avatar was able to perform actions including reach and gaze. In the Reach condition, the avatar was represented as a black circle with two curves above and below it, depicting its "body" and "tentacles". One "tentacle" was able to extend to the outer limit of the target following a correct response, implying its ability to *behave* (Fig. 3). The avatar occupied approximately 252 × 439 pixels (4.40° × 7.66°) and was positioned 190 pixels (3.32°) to the left or right of the central fixation cross. In the Gaze condition, the avatar was designed as a black circle with a smaller white circle inside, representing its "body" and "eye", but without "tentacles". The "pupil" of this avatar could shift 18 pixels (0.32°) to its left or right to align with the target, implying its ability to *perceive*. The avatar covered an area of about 252 × 252 pixels (4.40° × 4.40°), maintaining the same distance from the central fixation cross as the Reach condition (Fig. 3). In the Gaze&Reach condition, the avatar had both an "eye" and "tentacles", implying its ability to both *behave* and *perceive* (Fig. 3).

Avatar positions were manipulated by rotating the avatar by 90° and − 90°. Avatar placement was blocked, changing only after the first half of the experiment. Counterbalancing was implemented not only for the mapping of light and dark blue stimuli to the left ("q") and right ("p") responses and the initial positions of the avatar, but also for the sequence of Avatar's action among participants, using a Latin square design.

1.1.4. Procedure and design

The experiment employed a 3 (Avatar's action: Reach, Gaze, Gaze&Reach) × 2 (Compatibility: Incompatible, Compatible) within-subjects design. Pairings were deemed *compatible* when the target's location matched with the participant's response location (e.g., when the avatar was positioned to the *left* (90°) and participants pressed the left key for a light blue target, the target appeared to the *left* of the avatar). Conversely, mismatched pairings were labeled as *incompatible* (e.g., when the avatar was positioned to the *left* (90°) and participants pressed the left key for a light blue target, the target appeared to the *right* of the avatar).

The experiment was conducted in a dimly lit room. Participants were instructed to respond solely to the target color and to ignore any movements of the avatar. Specifically, participants were directed to place their left hand on the "q" button and their right hand on the "p" button. They were to press the left button ("q") when a light blue target appeared and the right button ("p") when a dark blue target appeared.

Each participant completed six blocks, each alternating between a left (90°) or right (−90°) avatar position, under three distinct Avatar's action conditions (Reach, Gaze, Gaze&Reach). Each block commenced with 20 practice trials, followed by 40 repetitions across four conditions, where either a dark blue or light blue target appeared to the left or right of the avatar (above or below the fixation cross). The order of trials was randomized within each block. Each block comprised 160 trials, culminating in a total of 960 trials per participant, which typically required 50–60 min to complete.

The fixation cross and avatar were displayed at the onset of each trial and remained visible throughout the experiment (Fig. 4). Targets were presented above or below the fixation after a 750-ms delay. Participants were then instructed to respond as quickly and accurately as possible. A correct response led to the avatar acting accordingly, such as reaching towards the target. Conversely, an incorrect response resulted in the avatar reaching or gazing towards an empty space, accompanied by a feedback tone. Additionally, responses that exceeded 1000 ms or were shorter than 100 ms were considered errors and also triggered a feedback tone. An interval of 1500 ms separated each response from the start of the subsequent trial.

1.2. Results

False responses, reaction times (RTs) longer than 1000 ms or shorter than 100 ms were removed from the analyses (see Böffel and Müsseler, 2019b for details). A total of 3.72 % trials were therefore excluded. We performed a 3 × 2 ANOVA with the within-subjects factors Avatar's action (Reach, Gaze, Gaze&Reach) and Compatibility (Compatible, Incompatible) and the dependent variable - mean RTs and error rate separately.[2] The Bonferroni method was employed for all post-hoc multiple comparisons.

1.2.1. Reaction times

The ANOVA analysis revealed that neither the main effect of Avatar's action ($F(2,70) = 0.69$, $p = .498$, $\eta_p^2 = 0.019$) nor Compatibility ($F(1,35) = 0.61$, $p = .440$, $\eta_p^2 = 0.017$) were significant. However, a significant interaction of the factors Compatibility and Avatar's action was observed, with $F(2,70) = 4.78$, $p = .015$, $\eta_p^2 = 0.220$ (Fig. 5). The analysis of simple effects showed that that only in the Gaze&Reach condition were the RTs in the Compatible condition ($M = 493$ ms, $SD = 70$) significantly faster than in the Incompatible condition ($M = 498$ ms, $SD = 71$), with a 5-ms advantage ($p = .007$), suggesting participants' spontaneous perspective taking of the avatar. But no compatible advantage was found in the separate Reach ($p = .702$) or Gaze condition ($p = .532$).

1.2.2. Error rate

Fig. 5 presents the mean error rates and standard deviations for each condition of Compatibility (compatible vs. incompatible) and Avatar's action (Reach, Gaze, Gaze&Reach). The error rates were as follows: Compatible condition: Reach ($M = 3.40$, $SD = 2.94$), Gaze ($M = 3.32$, $SD = 4.52$), Gaze&Reach ($M = 3.85$, $SD = 6.36$); Incompatible condition: Reach ($M = 3.75$, $SD = 3.86$), Gaze ($M = 3.94$, $SD = 4.55$), Gaze&Reach ($M = 4.05$, $SD = 6.43$).

A significant main effect of Compatibility was observed ($F(1,35) = 6.80$, $p = .013$, $\eta_p^2 = 0.163$). However, no significant main effect of Avatar's action ($F(2,70) = 0.15$, $p = .761$, $\eta_p^2 = 0.004$), or interaction between the two factors was found ($F(2,70) = 0.33$, $p = .705$, $\eta_p^2 = 0.009$).

Thus, the pattern of error rates suggested no speed-accuracy trade-offs.

1.3. Discussion

The results of Experiment 1 regarding RTs supported our hypothesis

[2] In addition to analyzing mean reaction times and error rates, we also examined logged RTs and performed logistic regression on response correctness due to the data's distributional characteristics (e.g., right-skewed RTs and the binomial nature of response correctness; Whelan, 2008; Van Breukelen, 2005). These additional analyses (see Supplementary Materials) produced consistent results across methods, except in Experiment 1, where logistic regression showed no significant compatibility effect. This difference suggests limitations in the original error rate analysis and underscores the value of employing multiple analytical approaches to better understand the data.

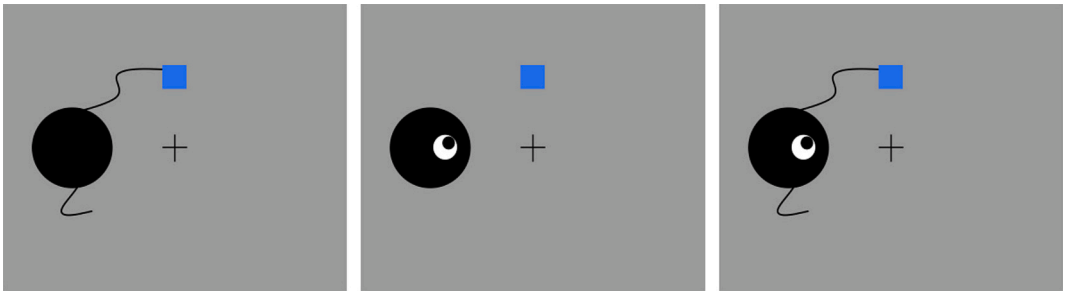

Fig. 3. The avatars used in Experiment 1.
Note. The figure shows schematic examples of the avatar's actions after a correct response in Experiment 1: (Left) Reach, (Middle) Gaze, and (Right) Gaze & Reach. Only the 90° rotated conditions, with the dark blue target positioned to the left of the avatar, are depicted.

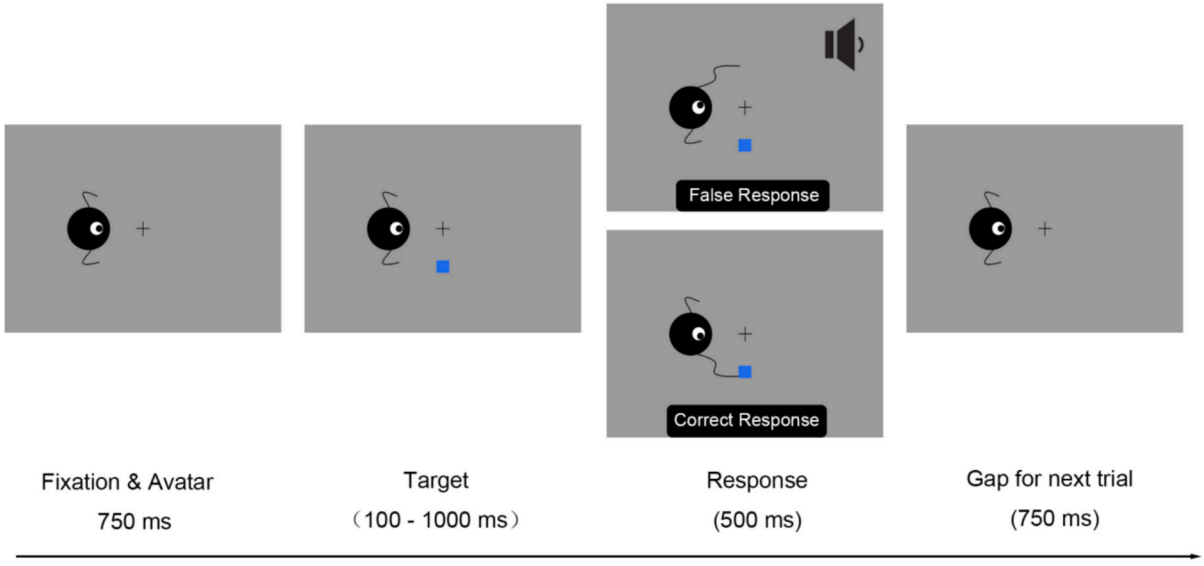

Fig. 4. Experimental procedure of Experiment 1.
Note. A schematic illustration of a single trial in Experiment 1. Only the avatar 90° rotated conditions, with the dark blue target positioned to the right of the avatar are depicted.

that a compatible advantage occurred solely under the Gaze&Reach condition. This suggests that even when the avatar did not resemble a human, participants would adopt its perspective when it simultaneously performed both gaze and reach. In contrast, no SR compatibility effect was observed in the single Reach or Gaze conditions, indicating that the avatar's possession of perceptual or behavioral abilities alone was insufficient to elicit spontaneous perspective-taking. While the results broadly aligned with our hypothesis, the compatible advantage under Gaze&Reach was small but significant, amounting to about 5 ms. Although the RT advantage in many prior SR compatible experiments was also modest, ranging from 3 ms to 40 ms (Böffel and Müsseler, 2019a; Böffel and Müsseler, 2020; Salm-Hoogstraeten and Müsseler, 2021), we aimed to replicate the effect in subsequent experiments to confirm the existence and robustness of this effect. Moreover, we observed that error rates in the compatible conditions were lower than in the incompatible conditions, suggesting that participants were generally less influenced by the avatar in the compatible condition. However, since the average error rate was quite low (less than 5 %) and there was no evidence of a speed-accuracy trade-off, we relied primarily on RTs as the main dependent variable for assessing the effect. This approach is consistent with common practice in previous studies (He et al., 2021; Samson et al., 2010; Wahn et al., 2023).

Before drawing conclusion, we need to exclude the potential confounding brought by the instruction to “ignore any movements of the avatar” in our experiments. This instruction might drive participants to focus on the avatar due to the rebound effect (Wegner, Schneider, Carter, & White, 1987), that is, suppressing thoughts does not inhibit the content that need to be suppressed, instead it leads to increased attention to them. For example, Wegner et al. (1987) demonstrated that when participants were told not to think about a white bear, the frequency of thinking about white bears actually increased. However, Böffel and Müsseler (2019b) ruled out the interference of this effect in the avatar-Simon paradigm. They found that whether participants were instructed to ignore the avatar (ignore condition) or adopt its perspective and imagine controlling its hands (steer condition), these instructions did not affect the resulting stimulus-response (SR) compatibility. To further confirm this in our study, we conducted a supplementary experiment with the same setup as Experiment 1, except for the instruction. In the supplementary experiment, participants were not explicitly instructed to ignore the avatar's movement. We compared the results of the original Experiment 1 (ignore condition) with Supplementary Experiment 1 (no ignore condition) and found no significant differences between the two groups (see Supplementary Materials). Therefore, we retained this instruction in subsequent experiments.

Moreover, while we have attempted to use simple shapes to create avatars performing actions, with the aim of reducing their human-like appearance, similar designs have been utilized in prior experiments, games, or cartoons as a metaphor for a human (Gao et al., 2010; Wu, 2015). This raises the possibility that participants may be familiar with the design, suggesting that the SR compatible effect may be influenced

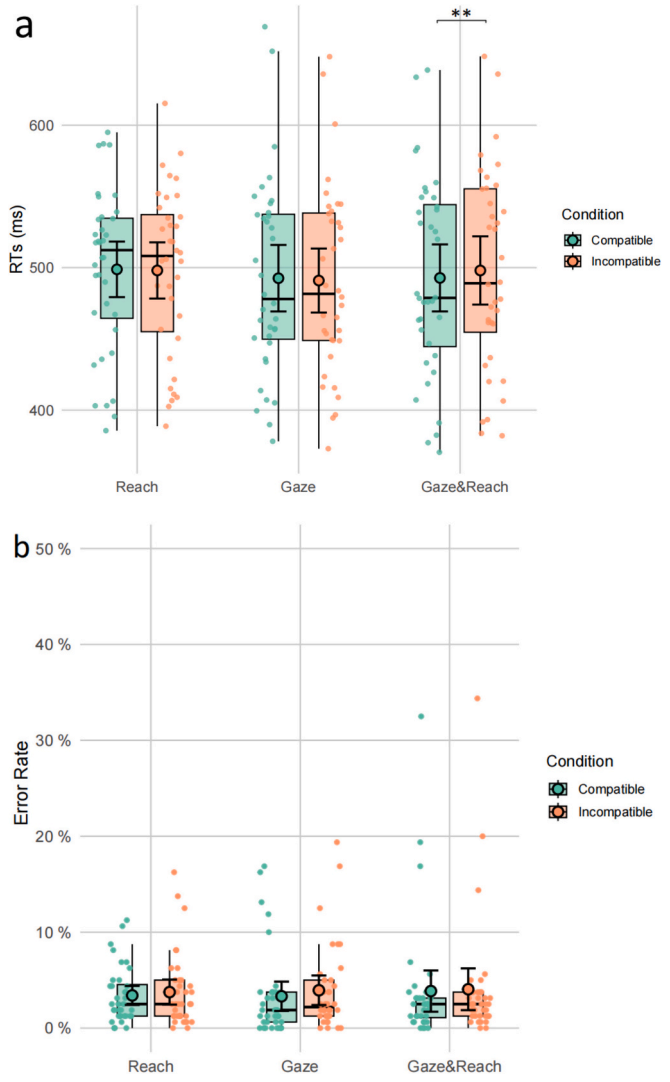

Fig. 5. Results of Experiment 1.
Note. The figure displays RTs (a) and error rates (b) in Experiment 1 across three conditions: Reach, Gaze, and Gaze&Reach, for both Compatible (green) and Incompatible (orange) trials. Asterisks represent a significant difference in RTs or error rate between the conditions (two for $p < .01$, three for $p < .001$). The boxplots show the median and interquartile range for each condition, with the whiskers extending to 1.5 times the interquartile range. The black circles indicate the mean RTs or error rates, with error bars representing 95 % confidence intervals. Individual data points are jittered around the boxplots to represent the distribution of RTs or error rates.

by an increased human-like appearance. To eliminate this possibility, a more mechanical appearance was employed in Experiment 2.

2. Experiment 2

Experiment 2 aimed to replicate the effects observed in Experiment 1 by employing avatars with a more mechanical appearance.

2.1. Method

2.1.1. Participants

Another 36 undergraduate students (20 females; age: $M = 19.89$, $SD = 2.05$) with normal or corrected-to-normal vision participated in the experiment for course credit or monetary compensation. All participants were provided informed consent and agreed to participate.

2.1.2. Apparatus, stimuli, and procedure

The experimental setup was similar to that of Experiment 1, with modifications only to the avatar's appearance (Fig. 6). The avatar was designed as a robot with two mechanical arms and one camera head. In the Reach condition, the avatar was represented as a black rectangle with two mechanical arms (264 × 406 pixels; 4.61° × 7.08°) but without the camera head. Upon a correct response, the avatar's arm extended towards the target's center, leaving a 49 pixels gap (0.86°), implying its ability to *behave*. In the Gaze condition, the avatar was a black rectangle equipped with a camera head (301 × 227 pixels; 5.25° × 3.97°) but without the mechanical arms. Correct responses caused the camera head to tilt 10 degrees towards the target implying its ability to *perceive*. In the Gaze&Reach condition, the avatar had both a camera head and two mechanical arms, implying its ability to both *behave* and *perceive*.

2.2. Results

False responses, RTs longer than 1000 ms or shorter than 100 ms were all regarded as errors and removed from the analyses (see Böffel and Müsseler, 2019a for details). A total of 3.93 % trials were excluded in this way. We performed a 3 × 2 ANOVA with the within-subjects factors Avatar's action (Reach, Gaze, Gaze&Reach) and Compatibility (Compatible, Incompatible) and the dependent variable - mean RTs and error rate separately. The Bonferroni method was employed for all post-hoc multiple comparisons.

2.2.1. Reaction times

The ANOVA analysis revealed a significant main effect of Compatibility ($F(1,35) = 28.05$, $p < .001$, $\eta_p^2 = 0.445$), indicating faster RTs in the Compatible condition than in the Incompatible condition. However, the main effect of Avatar's action was not significant ($F(2,70) = 1.75$, $p = .181$, $\eta_p^2 = 0.048$). A significant interaction between the factors Compatibility and Avatar's action was observed ($F(2,70) = 5.31$, $p = .010$, $\eta_p^2 = 0.132$) (Fig. 7). The analysis of simple effects showed that in the Gaze&Reach action, the Compatible condition ($M = 478$ ms, $SD = 69$) had significantly faster RTs than the Incompatible condition ($M = 489$ ms, $SD = 71$), with a 10-ms advantage ($p < .001$), The Gaze condition also showed a 5-ms advantage in the Compatible condition ($M = 473$ ms, $SD = 65$) compared to the Incompatible condition ($M = 478$ ms, $SD = 66$) ($p = .002$). No advantage was observed in the Reach condition ($p = .126$).

Additionally, we compared the compatible advantage between the Gaze&Reach and Gaze conditions. A paired-samples *t*-test revealed that the compatible advantage in the Gaze&Reach condition ($M = 10.44$, $SD = 8.82$) was significantly greater than that in the Gaze condition ($M = 5.47$, $SD = 9.62$), $t(35) = 2.74$, $p = .010$. This finding demonstrates that the spontaneous perspective-taking effect was stronger when the avatar performed both gaze and reach actions compared to gaze alone.

2.2.2. Error rate

Fig. 7 presents the mean error rates and standard deviations for each condition of Compatibility (compatible vs. incompatible) and Avatar's action (Reach, Gaze, Gaze&Reach). The error rates were as follows: Compatible condition: Reach ($M = 3.89$, $SD = 3.26$), Gaze ($M = 3.63$, $SD = 3.03$), Gaze&Reach ($M = 3.99$, $SD = 6.06$); Incompatible condition: Reach ($M = 3.73$, $SD = 3.04$), Gaze ($M = 3.79$, $SD = 2.75$), Gaze&Reach ($M = 4.53$, $SD = 4.85$).

There were no significant main effects for Compatibility ($F(1,35) = 0.36$, $p = .554$, $\eta_p^2 = 0.010$) or Avatar's action ($F(2,70) = 0.23$, $p = .794$, $\eta_p^2 = 0.013$), nor was there a significant interaction between the two factors ($F(2,70) = 1.08$, $p = .344$, $\eta_p^2 = 0.030$).

Thus, the pattern of errors suggested no speed-accuracy trade-offs.

2.3. Discussion

Experiment 2 essentially replicated the findings of Experiment 1, and

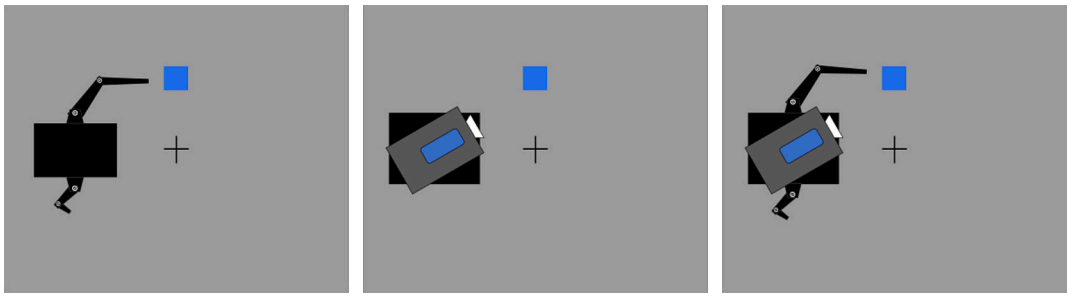

Fig. 6. The avatars used in Experiment 2.
Note. The figure illustrates schematic examples of the Avatar's action after correct response in Experiment 2: (Left) Reach. (Middle) Gaze. (Right) Gaze&Reach. Only the avatar 90° rotated conditions with the dark blue target positioned to the left of the avatar are depicted.

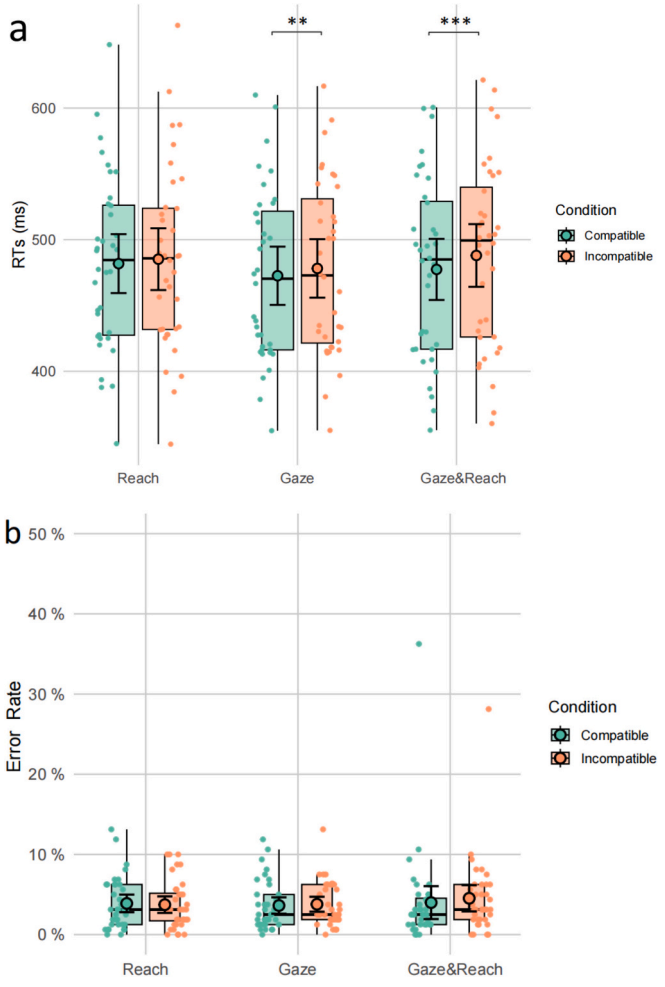

Fig. 7. Results of Experiment 2.
Note. The figure presents reaction times (a) and error rates (b) from Experiment 2 across three conditions: Reach, Gaze, and Gaze & Reach. Results are shown for both Compatible (green) and Incompatible (orange) trials. Plotting conventions match those used in Fig. 5.

generally demonstrated stronger effects (see Supplementary Materials for a comparison between Experiments 1 and 2). The 10-ms compatible advantage in the Gaze&Reach condition suggests that participants adopted the avatar's perspective despite its less human-like appearance. This spontaneous perspective-taking was also observed in the Gaze condition, which unexpectedly showed a 5-ms compatible advantage.

It should be noted, however, that comparisons between Experiments 1 and 2 revealed significantly larger effects in Experiment 2 across all conditions, but no significant interaction between Experiment and Avatar's action (see Supplementary Materials). This indicates that the significant finding in the Gaze condition may be attributed to these generally larger effects rather than a unique effect of the Gaze condition itself. We speculate that the camera's deflection in the Gaze condition may have implied a more pronounced behavioral response compared to the subtler pupil movement in Experiment 1. However, this interpretation lacks sufficient statistical evidence and warrants further investigation, which we will discuss more thoroughly in the general discussion. Nevertheless, the largest effect size was elicited under the combined Gaze&Reach condition, which aligned with our theoretical framework.

3. Experiment 3

Experiment 3 examined whether the causal order matters for the spontaneous perspective-taking triggered by the gaze and reach actions of non-human agents. We employed a 2 × 2 mixed design with a between-subjects factor of Order (Reach-Gaze, Gaze-Reach) and a within-subjects factor of Compatibility (Compatible, Incompatible).

3.1. Method

3.1.1. Participants

Another 48 undergraduate students (23 females; age: $M = 22.94$, $SD = 3.86$) with normal or corrected-to-normal vision participated in the experiment for course credit or monetary compensation. All participants were provided informed consent and agreed to participate.

G*Power 3.1 software was used to calculate the required sample size (Faul et al., 2007). Based on Experiment 2's findings, which had an actual power greater than 0.98 and an observed effect size (η_p^2) of 0.132 for the interaction between Avatar's action and Compatibility. The reported effect sizes (η_p^2) in this study include correlations between paired measures and were adjusted following the corrections proposed by Lakens (2013). As a result, a sample size of 46 participants was deemed sufficient. To ensure balanced conditions, the final sample size was increased to 48.

3.1.2. Apparatus, stimuli, and procedure

Experiment 3 retained only the Gaze&Reach action condition, divided into two sequences: Reach-Gaze and Gaze-Reach. All the settings, except the experimental procedure, remained consistent with Experiment 2. In the Gaze-Reach condition (Fig. 8a), following a correct response, the avatar first gazed for 500 ms, then performed both gaze and reach for another 500 ms before returning to fixation for 1500 ms until the next target appeared. This order is causally reasonable. The Reach-Gaze condition (Fig. 8b) followed a similar sequence but initiated with a reach, resulting in an order causally unreasonable. Each participant completed two blocks (avatar rotated 90° and − 90°), each containing 160 trials, totaling 320 trials with a duration of 20–30 min. The starting position was counterbalanced across participants.

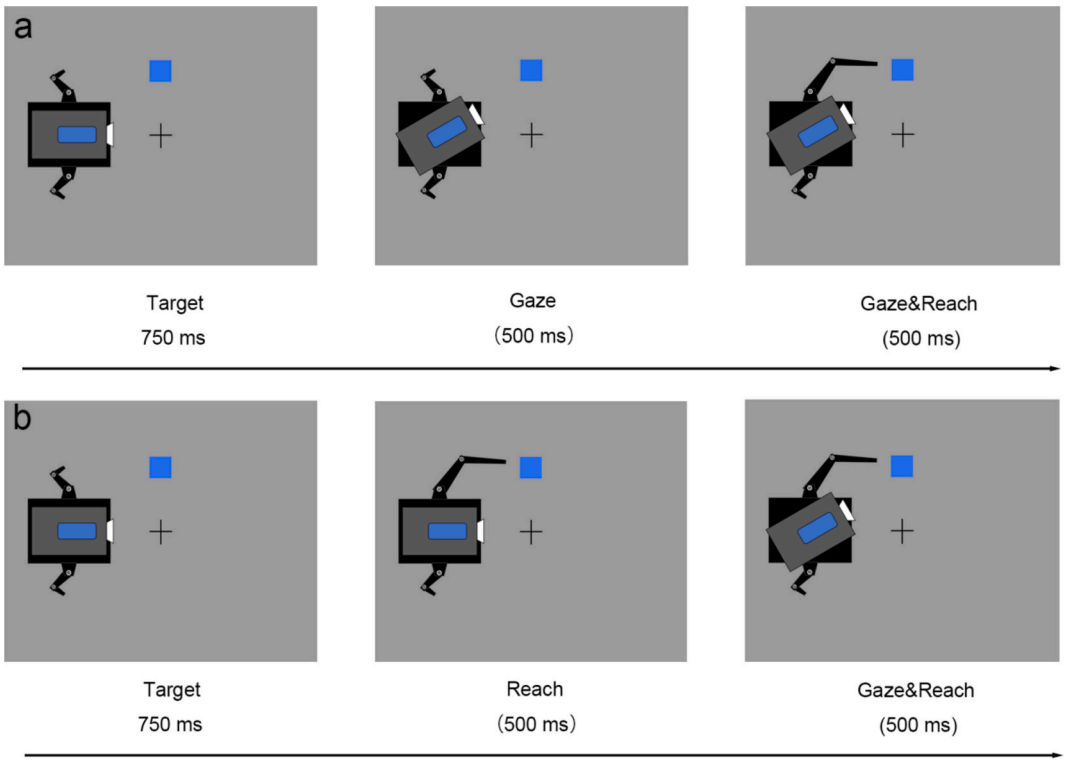

Fig. 8. The avatars used in Experiment 3.
Note. The figure illustrates schematic illustration of the avatar's actions procedure for the Gaze-Reach (a) and Reach-Gaze (b) condition in Experiment 3. Only the 90° avatar rotated conditions with the dark blue target positioned to the left of the avatar are depicted.

3.2. Results

False responses, RTs longer than 1000 ms or shorter than 100 ms were all regarded as errors and excluded from the analyses (see Böffel and Müsseler, 2019b for details). A total of 6.02 % of trials were excluded by this method. The mean RTs and error rate were analyzed separately using 2 × 2 mixed ANOVA with a within-subjects variable: Compatibility (Compatible, Incompatible), and a between-subjects variable: Order (Gaze-Reach, Reach-Gaze). Bonferroni correction was used for all post-hoc multiple comparisons.

3.2.1. Reaction times

The ANOVA analysis revealed a significant main effect of Compatibility ($F(1,46) = 5.30$, $p = .026$, $\eta_p^2 = 0.103$), indicating faster RTs in the Compatible condition than in the Incompatible condition. The main effect of Order was not significant ($F(1,46) = 0.67$, $p = .419$, $\eta_p^2 = 0.014$). However, a significant interaction between Order and Compatibility was observed ($F(1,46) = 11.98$, $p = .001$, $\eta_p^2 = 0.207$) (Fig. 9). Analysis of simple effects showed that in the Gaze-Reach condition, the Compatible condition ($M = 482$ ms, $SD = 79$) had significantly faster RTs than the Incompatible condition ($M = 492$ ms, $SD = 79$), with a 10-ms advantage ($p < .001$). No such advantage was found in the Reach-Gaze condition ($p = .417$).

3.2.2. Error rate

Fig. 9 presents the mean error rates and standard deviations for each condition of Compatibility (Compatible, Incompatible) and Order (Gaze-Reach, Reach-Gaze). The error rates were as follows: Compatible condition: Gaze-Reach ($M = 6.25$, $SD = 7.61$), Reach-Gaze ($M = 5.83$, $SD = 6.14$); Incompatible condition: Gaze-Reach ($M = 5.89$, $SD = 7.23$), Reach-Gaze ($M = 6.09$, $SD = 6.85$).

There was no significant main effects for Compatibility ($F(1,46) = 0.02$, $p = .900$, $\eta_p^2 = 0.000$), Order ($F(1,46) < 0.01$, $p = .958$, $\eta_p^2 = 0.000$), or interaction between the two factors ($F(1,46) = 0.58$, $p = .452$, $\eta_p^2 = 0.012$).

Thus, the pattern of errors suggested no speed-accuracy trade-offs.

3.3. Discussion

Experiment 3 supported the hypothesis that spontaneous perspective-taking occurs only when perception (gaze) precedes behavior (reach) causally. When the avatar executed gaze prior to the reach, participants spontaneously adopted the avatar's perspective. Conversely, when the reach was performed before the gaze, participants did not adopt the avatar's perspective.

4. Experiment 4a

This experiment was adapted from Samson et al.'s (2010) dot-counting task by introducing three conditions in which non-human avatars performed actions: Reach, Gaze, and Gaze & Reach. The aim was to generalize the effects observed in Experiments 1 and 2 to this well-established paradigm, which does not involve avatar actions or stimulus-response (SR) compatibility effects. The design effectively ruled out the possibility that the previously observed effects were driven by immersion or action control (Böffel and Müsseler, 2019a; von Salm-Hoogstraeten and Müsseler, 2021). In this adapted paradigm, avatar actions occurred *before* participants' key presses and were entirely unrelated to the task. This allowed us to test whether the observed effects in Experiments 1–3 were influenced by participants' sense of control over the avatar's actions, which in those earlier experiments occurred immediately *after* the participants' responses (Böffel and Müsseler, 2019b).

4.1. Method

4.1.1. Participants

The same sample size as in Experiments 1 and 2 was used, with 36

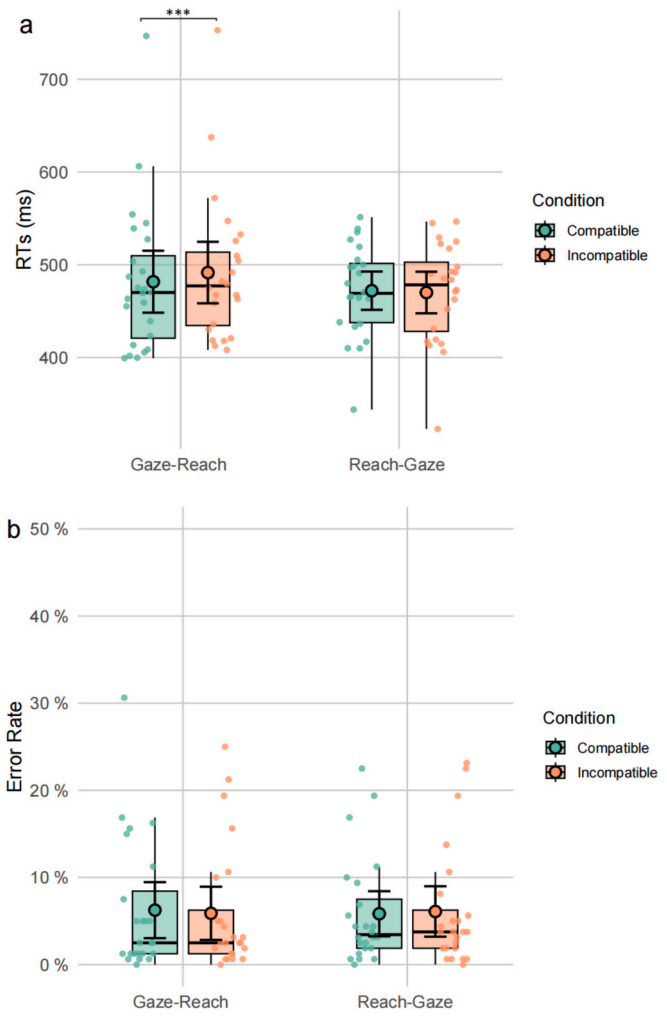

Fig. 9. Results of Experiment 3.
Note. The figure displays RTs (a) and error rate (b) in Experiment 3 across two conditions: Gaze-Reach, and Reach-Gaze, for both Compatible (green) and Incompatible (orange) trials. Plotting conventions match those used in Fig. 5.

undergraduate students (13 females; age: $M = 22.42$, $SD = 3.36$), all with normal or corrected-to-normal vision, participating for course credit or monetary compensation. Informed consent was obtained from all participants prior to the experiment.

4.1.2. Stimuli and procedure

According to the dot-counting task developed by Samson et al. (2010), the stimuli consisted of a picture showing a lateral view of a room with 0–3 red discs displayed on one or both walls. The avatar faced either the left or right wall. In 50 % of the trials (consistent condition), the number of discs visible from the avatar's perspective matched that from the participants' perspective (Fig. 1). In the other 50 % of trials (inconsistent condition), the number of discs differed between the two perspectives (Fig. 1). When asked to judge the number of discs from their own perspective, participants' performance decreased under inconsistent conditions, providing evidence of spontaneous perspective-taking, as the avatar's perspective was irrelevant to the task.

The avatar used in this experiment also had a non-human-like appearance, and was still capable of performing one or both human-like actions: Reach, Gaze, or Gaze&Reach (Fig. 10). In the Reach condition, the avatar appeared as a standing metal pillar equipped with extendable mechanical arms. Initially, the arms were lowered, but when a dot appeared on the wall, the avatar extended its arm towards the side that it faced. In the Gaze condition, the avatar featured a camera-like head mounted on the same metal pillar, but lacked an arm. Initially, the head was lowered, when a dot appeared, the avatar raised its head towards the wall which it faced. In the Gaze&Reach condition, the avatar combined both actions, raising its head and simultaneously reaching towards the wall that it faced.

Each trial began with a fixation cross displayed for 750 ms. After a 500-ms delay, the words "you" or "robot" appeared for 750 ms, indicating whether participants should adopt their own perspective or that of the avatar. Following another 500 ms, a digit (0–3) was shown for 750 ms, specifying the number of dots participants needed to verify from the indicated perspective. Then, the avatar, with its head lowered or arms hanging down and no dots on the walls, was presented for 500 ms. Afterward, the dots appeared on the walls, and the avatar simultaneously performed the reach, gaze, or gaze&reach actions. Participants were required to press one of two buttons ("f" for "yes" or "j" for "no") to judge whether the dots matched or mismatched the specified perspective. Feedback was provided for 750 ms, indicating whether the response was "correct" or "incorrect". If no response was given within 2000 ms, the response was labeled as incorrect, and the next trial began. (See Fig. 11).

Each participant needed to complete three blocks, with the avatar performing a specific action in each block. Each block consisted of 96 matching trials (where participants needed to press "yes"), including 48 consistent and 48 inconsistent conditions; and 96 mismatching trials (where participants needed to press "no"), also including 48 consistent and 48 inconsistent conditions. Additionally, there were 16 filler trials, in which no dots appeared on the walls. Participants completed 26 practice trials before each block. The stimuli within each block were presented in a random order. The order of the blocks was counterbalanced across participants using a Latin square design.

4.2. Results

Trials requiring a "no" response in the consistent condition were easier than in the inconsistent condition. Thus, only the trials requiring a "yes" response that was correctly answered were included in the data analysis (see Samson et al., 2010 for details). The dependent variable was the RTs and error rate for participants judging the number of discs only from their own perspective, as this condition directly related to the spontaneous perspective-taking effect we aimed to investigate. False responses, and RTs longer than 2000 ms were removed from the analyses (see Samson et al., 2010 for details). A total of 8.25 % of trials were therefore excluded. These RTs and error rates were analyzed using a 3 × 2 repeated-measures ANOVA, with Avatar's action (Reach, Gaze, Gaze&Reach) and perspective Consistency (Consistent vs. Inconsistent) as within-subject variables. Bonferroni correction was applied for all post-hoc multiple comparisons.

4.2.1. Reaction times

The ANOVA analysis revealed a significant main effect of Consistency ($F(1,35) = 106.07$, $p < .001$, $\eta_p^2 = 0.752$), indicating faster RTs in the Consistent condition than in the Inconsistent condition. However, the main effect of Avatar's action was not significant ($F(2,70) = 1.91$, $p = .160$, $\eta_p^2 = 0.052$). A significant interaction between the factors Consistent and Avatar's action was observed ($F(2,70) = 20.92$, $p < .001$, $\eta_p^2 = 0.374$) (Fig. 12a). The analysis of simple effects showed that in the Gaze&Reach action, the Consistent condition ($M = 434$ ms, $SD = 147$) had significantly faster RTs than the Inconsistent condition ($M = 564$ ms, $SD = 167$), with a 130-ms advantage ($p < .001$), The Gaze condition also showed a 105-ms advantage in the Consistent condition ($M = 428$ ms, $SD = 166$) compared to the Inconsistent condition ($M = 533$ ms, $SD = 163$) ($p < .001$). We further conducted a paired t-test on the consistent advantage between Gaze and Gaze&Reach, and found that the consistent advantage on Gaze&Reach condition was significantly larger than that under Gaze condition ($t(35) = 2.11$, $p = .042$, Cohen's $d = 0.714$). No advantage was observed in the Reach condition ($p = .105$).

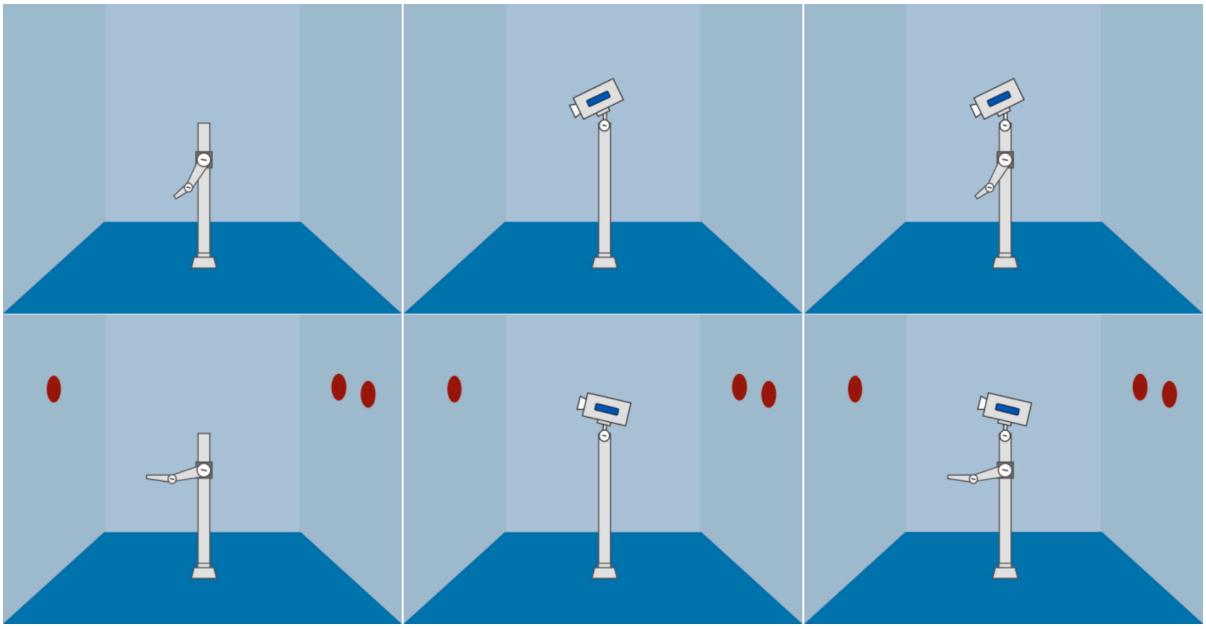

Fig. 10. The schematic illustration of Experiment 4a.
Note. The figure illustrates the three Avatar's action conditions (Reach, Gaze, Gaze&Reach) from left to right in Experiment 4a. The top row shows the avatar figure in each condition before the dots appeared, while the bottom row displays the avatar's corresponding action when the dots appeared.

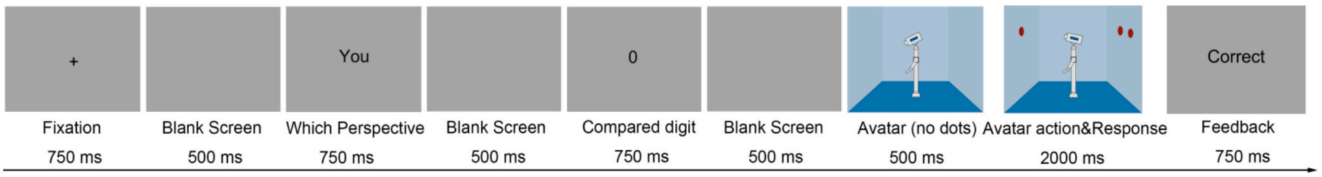

Fig. 11. The schematic procedure of Experiment 4a.
Note. The figure illustrates the procedure of a trial in Experiment 4a.

4.2.2. Error rate

Fig. 12b presents the mean error rates and standard deviations for each condition of Consistency (consistent vs. inconsistent) and Avatar's action (Reach, Gaze, Gaze&Reach). The error rates were as follows: Consistent condition: Reach ($M = 3.24$, SD = 5.09), Gaze ($M = 2.66$, $SD = 4.12$), Gaze&Reach ($M = 3.94$, $SD = 6.68$); Inconsistent condition: Reach ($M = 13.77$, $SD = 12.23$), Gaze ($M = 11.81$, $SD = 8.42$), Gaze&Reach ($M = 14.10$, $SD = 10.11$).

There was a significant main effect for Consistency ($F(1,35) = 81.84$, $p < .001$, $\eta_p^2 = 0.700$, but no significant effect for Avatar's action ($F(2,70) = 1.63$, $p = .208$, $\eta_p^2 = 0.044$), or interaction between the two factors ($F(2,70) = 0.29$, $p = .733$, $\eta_p^2 = 0.008$).

Thus, the pattern of errors suggested no speed-accuracy trade-offs.

4.3. Discussion

This result essentially replicated our findings from Experiment 2: when the avatar performed both gaze and reach actions, or gaze alone, spontaneous perspective-taking was elicited, with the effect being stronger in the Gaze&Reach condition. However, the reach-only condition did not produce any effect. For this result, we propose two possible accounts. First, gaze may play a more critical role than reach in eliciting perspective-taking. This could be due to its precedence in the causal sequence—perception typically occurs first in cognitive processing, and it is reasonable to perceive information without responding. Second, as noted in Experiment 2, the camera head movement in the gaze condition could subtly imply behavior, as the action to seek further information is also kind of a consequence of initial perception. To investigate this further, we plan to remove the camera movement in the gaze condition and assess whether spontaneous perspective-taking still occurs.

5. Experiment 4b

This experiment aimed to further reduce potential behavior-related cues associated with gaze. In both the Gaze and Gaze&Reach conditions, we removed the process of raising the robot's head. From the moment before the dot appeared to after it appeared, the robot's gaze remained towards the wall. Our objective was to observe whether the perspective-taking effect would persist under these modified Gaze and Gaze&Reach conditions.

5.1. Method

5.1.1. Participants

Another 36 undergraduate students (19 females; age: $M = 22.83$, $SD = 2.89$) with normal or corrected-to-normal vision participated in the experiment for course credit or monetary compensation. All participants were provided informed consent and agreed to participate.

5.1.2. Stimuli and procedure

The experimental setup was identical to that of Experiment 4a, except the modifications made to the avatar's action: the robots maintained their gaze towards the wall instead of lowering their heads before the dots appeared in the Gaze and Gaze&Reach conditions (Fig. 13).

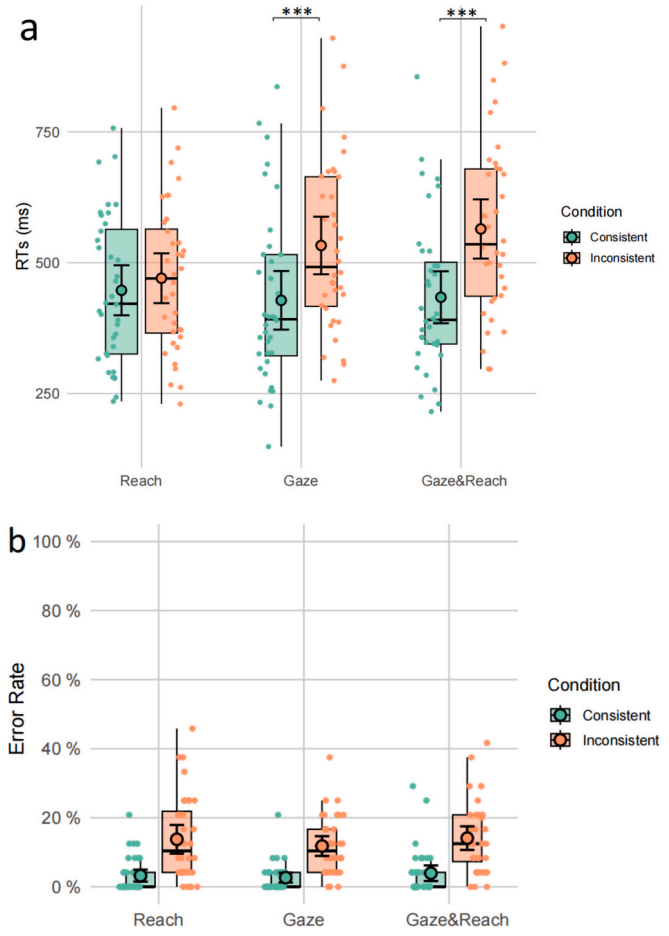

Fig. 12. Results of Experiment 4a.
Note. The figure displays RTs (a) and error rate (b) in Experiment 4a across three conditions: Reach, Gaze, and Gaze&Reach, for both Consistent (green) and Inconsistent (orange) trials. Plotting conventions match those used in Fig. 5.

5.2. Results

False responses, RTs longer than 2000 ms were removed from the analyses (see Samson et al., 2010 for details). A total of 9.10 % trials were therefore excluded. As in Experiment 4a, we only considered the "yes" responses. We analyzed these RTs using a 3 × 2 repeated-measures ANOVA, with avatar's action (Reach, Gaze, Gaze&Reach) and perspective Consistency (Consistent, Inconsistent) as within-subject variables. Bonferroni correction was applied for all post-hoc multiple comparisons.

5.2.1. Reaction times

The ANOVA analysis revealed a significant main effect of Consistency ($F(1,35) = 58.94$, $p < .001$, $\eta_p^2 = 0.627$), indicating faster RTs in the Consistent condition than in the Inconsistent condition. However, the main effect of Avatar's action was not significant ($F(2,70) = 0.25$, $p = .768$, $\eta_p^2 = 0.007$). A significant interaction between the factors Consistent and Avatar's action was observed ($F(2,70) = 22.94$, $p < .001$, $\eta_p^2 = 0.396$) (Fig. 14a). The analysis of simple effects showed that in the Gaze&Reach action, the Consistent condition ($M = 455$ ms, $SD = 164$) had significantly faster RTs than the Inconsistent condition ($M = 630$ ms, $SD = 190$), with a 175-ms advantage ($p < .001$), The Gaze condition also showed a 127-ms advantage in the Consistent condition ($M = 466$ ms, $SD = 165$) compared to the Inconsistent condition ($M = 593$ ms, $SD = 187$) ($p < .001$). We further conducted a paired t-test on the consistent advantage between Gaze and Gaze&Reach, and found that the consistent advantage on Gaze&Reach condition was significantly larger than that under Gaze condition ($t(35) = 2.79$, $p = .009$, Cohen's $d = 0.943$). No advantage was observed in the Reach condition ($p = .216$).

5.2.2. Error rate

Fig. 14b presents the mean error rates and standard deviations for each condition of Consistency (Consistent vs. Inconsistent) and Avatar's action (Reach, Gaze, Gaze&Reach). The error rates were as follows: Consistent condition: Reach ($M = 4.05$, $SD = 5.13$), Gaze ($M = 3.13$, $SD = 4.15$), Gaze&Reach ($M = 3.13$, $SD = 3.78$); Inconsistent condition: Reach ($M = 15.16$, $SD = 10.92$), Gaze ($M = 13.54$, $SD = 11.37$), Gaze&Reach ($M = 15.63$, $SD = 10.51$) (Fig. 14b).

There was a significant main effect for Consistency ($F(1,35) = 136.20$, $p < .001$, $\eta_p^2 = 0.796$, but no significant effect for Avatar's action ($F(2,70) = 0.60$, $p = .549$, $\eta_p^2 = 0.017$), or interaction between the two factors ($F(2,70) = 20.98$, $p = .713$, $\eta_p^2 = 0.009$).

Thus, the pattern of errors suggested no speed-accuracy trade-offs.

5.3. Discussion

This result once again replicated the result of Experiment 4a: when the avatar performed both gaze and reach actions, or gaze alone, spontaneous perspective-taking was elicited, with a stronger effect in the Gaze&Reach condition. However, the Reach-only condition did not produce any effect. These findings suggest that gaze, as a perceptual action, may play a more pivotal role than behavioral actions alone in triggering perspective-taking. This could be due to its precedence in the causal sequence or its inherent implication of a potential behavioral response. We will discuss this point further in the general discussion.

6. General discussion

6.1. Two key factors for spontaneous perspective-taking occurring on non-human agents

Contradictory evidence exists regarding the role of human-like appearance in eliciting spontaneous perspective-taking. While some studies report that agents closely resembling humans can elicit this phenomenon, others do not find a significant effect (Furlanetto et al., 2013; Salm-Hoogstraeten and Müsseler, 2021; Wahn and Berio, 2023; Xiao et al., 2022; Ye et al., 2023). Our hypothesis challenges the view that only an agent "looks" like a human is crucial. Instead, whether it "acts" like a human — specifically perceptual and behavioral abilities — is also critical. Our research focused on two goal-directed actions, gaze and reach, and found that agents without a human-like appearance could elicit spontaneous perspective-taking if they performed both actions. Conversely, agents limited to a single action did not trigger perspective-taking (Experiment 1). This effect was consistent even when the agent presented a distinctly non-human, mechanical appearance (Experiment 2). Furthermore, our results underscore the importance of the causal sequence of these actions. Participants spontaneously adopted the agent's perspective only when gaze preceded reach (Experiment 3). We further extended our findings from Experiment 1&2 by employing another traditional perspective-taking paradigm—the dot-counting task designed by Samson et al. (2010) (Experiment 4). Therefore, our studies suggest that spontaneous perspective-taking depends not simply on an agent's appearance but on its ability to engage in a human-like interactive pattern.

The stronger effects of spontaneous perspective-taking in Experiment 2 may be due to the avatar's more complex and larger design compared to Experiment 1.[3] In Experiment 2, the avatar's larger mechanical arms and distinct rotating camera-like head made the actions of gazing and reaching more visually salient than the slender tentacle and subtle pupil

[3] We are grateful to an anonymous reviewer for raising the issue discussed in this and the following paragraph.

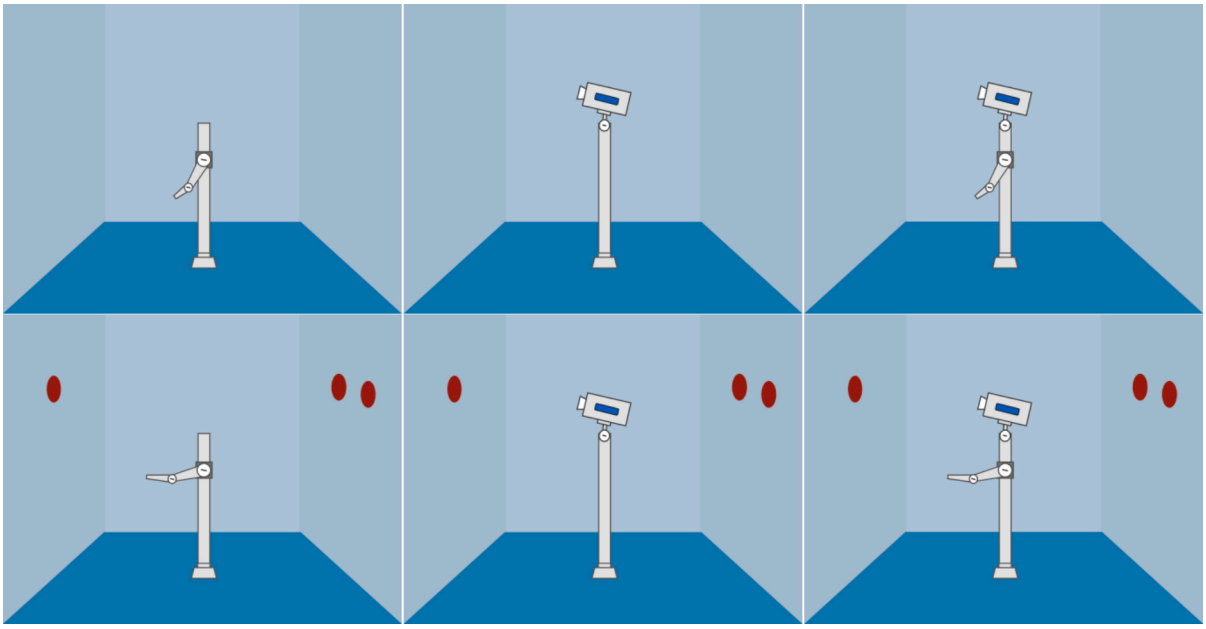

Fig. 13. The schematic illustration of Experiment 4b.
Note. The figure illustrates the three Avatar's action conditions (Reach, Gaze, Gaze&Reach) from left to right in Experiment 4b. The top row shows the avatar figure in each condition before the dots appeared, while the bottom row displays the avatar's corresponding action when the dots appeared.

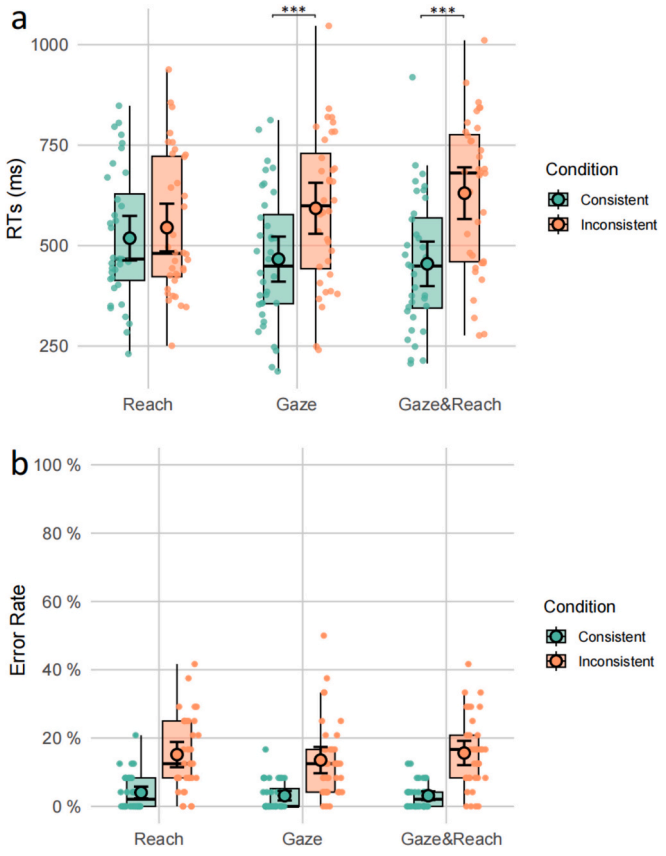

Fig. 14. Results of Experiment 4b.
Note. The figure displays RTs (a) and error rate (b) in Experiment 4b across three conditions: Reach, Gaze, and Gaze&Reach, for both Consistent (green) and Inconsistent (orange) trials. Plotting conventions match those used in Fig. 5.

movements in Experiment 1. This increased action salience likely captured more attention from participants (Rutkowska et al., 2022; Yantis and Egeth, 1999), thereby enhancing the strength of the observed effect. Moreover, the absence of a significant interaction between Experiment and Avatar's action (Appendix C) suggests that the differences in effect strength could be linked to the overall increased visual prominence of the avatar's movements. While we speculated above that features such as camera head movements might simultaneously convey both perceptual and behavioral cues, our current analysis lacked strong statistical evidence to confirm this interpretation. Future research could systematically investigate how these design elements influence spontaneous perspective-taking by determining whether and how such visual cues communicate the avatar's perceptual and behavioral abilities.

In addition, our result suggested that gaze is more important than reach for triggering spontaneous perspective taking, as Experiment 2 showed an effect in the gaze-only condition, and in Experiment 3 the effect persisted only when gaze preceded reach. Similar results were also found in Experiment 4. We propose two possible explanations for this finding. Firstly, from an information-processing perspective, perception is a prerequisite for behavior, while perception does not necessarily lead to behavioral response. According to the perceived information, one can decide whether or not to perform an act, therefore perception does not inherently follow with an overt behavior. Secondly, we propose that gaze itself may convey implications of behavioral response. Since others' intentions cannot be directly read, we often infer them through behaviors (Dennett, 1989). Gazing, as a perceptual action performed to gather information, indicating that the agent is gathering information, similar to the action that people use eye direction to express intention (Castiello, 2003), which is also a kind of behavioral outcome. Additionally, the avatar's "camera head" that enables it to "see" could also metaphor a behavioral outcome without overt action—namely, transmitting the captured images to the security behind the monitor screen, which is the function that cameras are supposed to have. This may potentially make gaze more significant than reach in certain contexts.

An alternative explanation for the results may be action-effect integration. In Experiment 1 and 2, the avatar's reach and gaze actions occurred immediately following participants' correct responses, which may have introduced action effects influencing the results. According to

event coding theory (Hommel et al., 2001), when actions and their effects overlap in time, action-effect integration occurs. Participants likely learned the relationship between their responses (actions) and the avatar's actions (effects), enhancing their sense of control over the avatar's behavior and thereby strengthening the stimulus-response (SR) compatibility effect (Böffel and Müsseler, 2019a). However, we propose that this does not significantly impact the interpretation of our results. Specifically, we observed that when the non-human avatar performed only the reaching action, the SR compatibility effect did not emerge (Experiments 1&2). In contrast, Böffel and Müsseler (2019b) found that when a human avatar performed reaching, it produced a stronger effect. This difference is likely because the human avatar inherently implied perceptual and behavioral capabilities. In our experiment, the non-human avatar, which lacked perceptual features, could not produce SR compatibility even with sufficient action feedback (reach). To further confirm our theory, we replicated the results in Experiment 4, which used a traditional perspective-taking paradigm—the dot-counting task designed by Samson et al. (2010). In this design, action effects are less likely to interfere, as the avatar displayed actions before the participant's response. The results indicated that the compatibility effect did not appear in the reach-only condition but was present in the gaze-only and combined gaze and reach conditions. Further analysis showed that the effect was significantly stronger in the Gaze&Reach condition compared to the gaze-only condition. Therefore, we argue that the avatar's human-like actions play a more pivotal role in producing these results.

Our findings offer insights into reconciling existing contradictory evidence in the field of spontaneous perspective-taking. For example, Xiao et al. (2022) reported that a robot avatar with a highly human-like appearance did not elicit spontaneous perspective-taking among participants. This may be because the avatars in Xiao et al.'s study featured uniform coloration across the head and face, likely obscuring participants' ability to discern the avatar's visual focus. Similarly, in experiments by Wahn and Berio (2023), a robot avatar with a less human-like, more animal-like appearance also failed to trigger this response. However, when this animal-like avatar was equipped with a camera head, participants readily adopted its perspective. This was likely due to the animal-like avatars used by Wahn and Berio (2023) having eyes colored black and yellow, which may have confused participants about the avatars' ability to perceive their surroundings. In contrast, our study presented a robot with a movable camera head, which provided a clear indication of its perceptual ability as well as the current focus of attention. Such a dynamic demonstration, rather than the static image that ambiguously implies the perceptual ability, could be more likely to trigger spontaneous perspective-taking. Based on our findings, these discrepancies are caused by the avatars' clarity in expressing perceptual abilities. Additionally, the perceptual indication of attention should be dynamic like the camera head gazing towards the target in the current experiments, rather than static images used in past studies, which often made it difficult to distinguish between an inanimate figure and a “live” robot capable of action.

Furthermore, our results underscore the socially adaptive nature of perspective-taking. Historically, perspective-taking was viewed as a stimulus-driven process, involving reflective response to gaze cues, shifting attention, and subsequently taking perspectives (Samson et al., 2010; Santiesteban et al., 2014). Recent evidence challenges this view, indicating that perspective-taking is not simply bottom-up but inherently social, accounting for the agent's potential for interaction (O'Grady et al., 2020; Zhou et al., 2022). For example, perspective-taking does not occur when an agent's ability to interact is obstructed by blocked sightlines (Baker et al., 2016; Cole et al., 2016; Conway et al., 2017; Furlanetto et al., 2016; O'Grady et al., 2020). It is aligned with our findings that perspective-taking is triggered not by the physical appearance of a stimulus alone, but by the perceived social interactive potential, specifically perceptual and behavioral abilities inferred from visual cues.

6.2. How to act like a human: behavior is the consequence of perception

Based on the findings above, to make an agent act human-like, the agent's behavior should be a consequence of its perception, with a reasonable causal relationship. When this causality is maintained, agents are perceived as potential human collaborators, and individuals are more likely to adopt their perspectives. Conversely, when this causality is disrupted, such as when an agent behaves randomly, it is no longer treated as human-like, and perspective-taking is not engaged. For instance, Zhao et al. (2015a) found that participants adopted the perspective of a robot equipped with eyes and capable of reaching. However, if the robot lacked eyes but could still reach, participants did not adopt its perspective. This finding supports our theory that behavior without preceding perception disrupts the necessary causal link, thereby does not lead to perspective-taking.

Additionally, the significance of causality extends beyond robots to human avatars. In experiments by von Salm-Hoogstraeten and Müsseler (2021), when a human avatar reached randomly, without considering participants' responses, people were less inclined to adopt its perspective. This pattern is also evident in other areas of social cognition, such as gaze-following behaviors. Abubshait and Wiese (2017) demonstrated that even when interacting with humans, if the agents randomly cued the location of an upcoming target, participants were less likely to follow their gaze direction. These findings further underscore the critical role of causality in fostering human-like interactive patterns essential for social cognition.

The causality we refer to here is the perceived causality, rather than simply the temporal order of events: causes precede effects. In the Gaze&Reach conditions of Experiments 1 and 2, while gaze and reach actions occur simultaneously; however, from a human observer's perspective, perception still precedes action. This is because perception usually starts before the observable behavioral changes (such as gaze shifts). The process usually starts with locating a potential target by peripheral vision, and then attention shifts to the target to guide saccadic, which are then followed by overt gaze shifts to gather detailed information and initiate a behavioral response (Deubel and Schneider, 1996; Henderson et al., 1989; Zhao et al., 2012). To an external observer, it may seem that the gaze shift and behavioral response happen almost at the same time, but pre-attention/gaze processing has been performed to gather part of the information for action. Therefore, gazing and reaching occurring simultaneously does not violate the principle that perception precedes behavior. Take a baseball game for an example, a player's head and body may rotate almost simultaneously, but this movement is pre-planned, allowing the player to respond immediately after perceiving the ball's position.

Compared to the temporal sequence of events, it is more crucial that the behavioral response must align with the perceived information. In the Gaze&Reach condition in our Experiments 1 and 2, the reaching and gazing actions are always aligned (towards the same target). However, if they do not align, it obviously defies typical human action patterns. For instance, if a baseball player were to focus visually on one ball but successfully hit a different one without turning their head, it would defy common sense. This is consistent with our finding that when the causal relationship was violated, the effect disappeared (Experiment 3).

6.3. Implications for human-robot interaction

Our findings also have implications for the design of robots aimed at enhancing human interaction. Firstly, though a robot's appearance is unnecessary to be highly human-like, it must be designed with a distinct perceptual system that clearly indicates it is sensing the external environment and specifies what it is sensing. We can find many current practices in robotics engineering ignore the explicit expression of sensing, for instance, many robots or autonomous vehicles are equipped with hidden cameras or laser radar (Singh et al., 2022), leading to uncertainty about their focus, which can be disconcerting for humans.

More user-friendly design requires a clear indicator of the sensing behavior, even simply adding a signal light that synchronizes with the cameras or radar could be highly beneficial to improving the usability in social interaction.

Secondly, the design of a robot should clearly convey its specific social interactive potential from the outset, prominently showcasing both the robot's capabilities and its limitations, rather than incorporating imperceptible elements that may confuse users. For example, from R2-D2's appearance, it is difficult for users firstly encountering it to imagine its ability to toss or grasp objects. However, those familiar with it understand that it actually has an internal tossing mechanism and mechanical arms. Therefore, an optimal design should incorporate more obvious features, such as visible springs and externally located mechanical arms. This would provide clear visual cues about what the robot can do, reducing confusion and facilitating more intuitive interaction.

Thirdly, and most importantly, the robot must demonstrate a perceivable perception-behavior causality to establish a common understanding with humans. Perception should precede behavior, with a human-perceivable time interval between the perceptual cause and the behavioral consequence, even if the robot does not require such a delay. Both its perceptual and behavioral pattern should align with rational principles understandable to humans, and the perceptual information required for behavior must be visibly obtained beforehand. For example, a robot equipped with 360° radar can quickly perceive and respond to objects behind it. However, since humans typically cannot see or react to things outside their line of sight, they might find this behavior confusing. Therefore, when designing robots, it is important to include a scanning light to clearly indicate that the robot has detected something behind it and is responding accordingly. This helps to create a clear and understandable causality that aligns with human perceptual and behavioral expectations.

6.4. Conclusion

The current study showed that, when the agent with which people engaged lacked a human-like appearance, spontaneous perspective-taking was observed if it demonstrated the ability to both gaze and reach, with reach contingent upon prior gaze. The results indicate that even when the agent is highly non-human-like, if it possesses both perceptual and behavioral abilities and follows a reasonable relationship where behavior is a consequence of perception, people will spontaneously adopt the agent's perspective. The findings suggest that robot designs should clearly indicate these abilities and causality (e.g., salient scanning light and perceptible pauses between scanning and action) to be better understood by humans and to facilitate better cooperation.

6.5. Constraints on generality

The present study identifies two essential human-like actions that trigger spontaneous perspective-taking in non-human agents. These findings were obtained from samples at Zhejiang University, consisting mainly of undergraduate and graduate students. While we did not exclusively use WEIRD (Western, Educated, Industrialized, Rich, and Democratic) populations, we recognize that research involving participants from educated, industrialized societies may not necessarily generalize to all populations. This limitation warrants careful consideration, and future research should aim to include more diverse samples to validate the universality of these findings. Given that perspective-taking is a fundamental social cognitive ability studied across various paradigms and populations, we expect the results to be applicable to adult populations. Additionally, since all avatars used in these experiments are 2D images, which differ somewhat from actual robots, we hope our results can be generalized to virtual non-human figures, such as characters in films or avatars controlled by players in video games. It is also important to note that factors affecting social cognition, such as personality traits or neurodiverse conditions (e.g., autism, alexithymia), might influence the observed effects.

CRediT authorship contribution statement

Xucong Hu: Writing – original draft, Software, Methodology, Formal analysis. **Haokui Xu:** Writing – review & editing, Visualization, Data curation. **Hui Chen:** Resources, Investigation. **Mowei Shen:** Funding acquisition. **Jifan Zhou:** Writing – review & editing, Supervision, Project administration, Conceptualization.

Declaration of competing interest

Authors have no conflicts of interest to disclose.

Acknowledgement

This research was supported by the National Natural Science Foundation of China (Grants 62337001, 32371088), Science and Technology Innovation 2030—"Brain Science and Brain-like Research" Major Project (2022ZD0210800), and the Fundamental Research Funds for the Central Universities (226-2024-00118).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2025.106101.

Data availability

All stimuli and data have been made available at the Open Science Framework (OSF) and can be accessed at https://osf.io/3dypm/.

References

Abrini, M., Auvray, M., & Chetouani, M. (2023, August). Humans' spatial perspective-taking when interacting with a robotic arm. In *2023 32nd IEEE international conference on robot and human interactive communication (RO-MAN)* (pp. 533–540). IEEE.

Abubshait, A., & Wiese, E. (2017). You look human, but act like a machine: Agent appearance and behavior modulate different aspects of human–robot interaction. *Frontiers in Psychology, 8*, 1393.

Baker, L. J., Levin, D. T., & Saylor, M. M. (2016). The extent of default visual perspective taking in complex layouts. *Journal of Experimental Psychology: Human Perception and Performance, 42*(4), 508–516.

Böffel, C., & Müsseler, J. (2018). Perceived ownership of avatars influences visual perspective taking. *Frontiers in Psychology, 9*, 743.

Böffel, C., & Müsseler, J. (2019a). Action effect consistency and body ownership in the avatar-Simon task. *PLoS One, 14*(8), Article e0220817.

Böffel, C., & Müsseler, J. (2019b). Visual perspective taking for avatars in a Simon task. *Attention, Perception, & Psychophysics, 81*, 158–172.

Böffel, C., & Müsseler, J. (2020). No evidence for automatic response activation with target onset in the avatar-compatibility task. *Memory & Cognition, 48*, 1249–1262.

Carlson, L., Skubic, M., Miller, J., Huo, Z., & Alexenko, T. (2014). Strategies for human-driven robot comprehension of spatial descriptions by older adults in a robot fetch task. *Topics in Cognitive Science, 6*(3), 513–533.

Castiello, U. (2003). Understanding other people's actions: Intention and attention. *Journal of Experimental Psychology: Human Perception and Performance, 29*(2), 416.

Clements-Stephens, A. M., Vasiljevic, K., Murray, A. J., & Shelton, A. L. (2013). The role of potential agents in making spatial perspective taking social. *Frontiers in Human Neuroscience, 7*, 497.

Cole, G., Atkinson, M., Le, A. T., & Smith, D. T. (2016). Do humans spontaneously take the perspective of others? *Acta Psychologica, 164*, 165–168.

Conway, J. R., Lee, D., Ojaghi, M., Catmur, C., & Bird, G. (2017). Submentalizing or mentalizing in a level 1 perspective-taking task: A cloak and goggles test. *Journal of Experimental Psychology: Human Perception and Performance, 43*(3), 454–465.

Dennett, D. C. (1989). *The intentional stance*. MIT press.

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research, 36*(12), 1827–1837.

Dolk, T., Hommel, B., Prinz, W., & Liepelt, R. (2013). The (not so) social Simon effect: A referential coding account. *Journal of Experimental Psychology: Human Perception and Performance, 39*(5), 1248.

Emery, N. J. (2000). The eyes have it: The neuroethology, function and evolution of social gaze. *Neuroscience & Biobehavioral Reviews, 24*(6), 581–604.

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. *Behavior Research Methods, 39*(2), 175–191.
Flavell, J. H. (1977). *Cognitive development*. Prentice-Hall.
Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young children's knowledge about visual perception: Further evidence for the level 1–level 2 distinction. *Developmental Psychology, 17*(1), 99.
Freina, L., Bottino, R., Tavella, M., & Chiorri, C. (2017). Evaluation of visuo-spatial perspective taking skills using a digital game with different levels of immersion. *International Journal of Serious Games, 4*(3).
Freundlieb, M., Kovács, Á. M., & Sebanz, N. (2016). When do humans spontaneously adopt another's visuospatial perspective? *Journal of Experimental Psychology: Human Perception and Performance, 42*(3), 401.
Freundlieb, M., Sebanz, N., & Kovács, Á. M. (2017). Out of your sight, out of my mind: Knowledge about another person's visual access modulates spontaneous visuospatial perspective-taking. *Journal of Experimental Psychology: Human Perception and Performance, 43*(6), 1065.
Furlanetto, T., Cavallo, A., Manera, V., Tversky, B., & Becchio, C. (2013). Through your eyes: Incongruence of gaze and action increases spontaneous perspective taking. *Frontiers in Human Neuroscience, 7*, 455.
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. *Journal of Experimental Psychology: Human Perception and Performance, 42*(2), 158.
Gao, T., McCarthy, G., & Scholl, B. J. (2010). The wolfpack effect: Perception of animacy irresistibly influences interactive behavior. *Psychological Science, 21*(12), 1845–1853.
Garofalo, G., Gawryszewski, L. L., & Riggio, L. (2022). Seeing through the cat's eyes: Evidence of a spontaneous perspective taking process using a non-human avatar. *Cognitive Processing, 23*(2), 269–283.
Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition, 125*(1), 125–130.
Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*(5812), 619.
Guttman, N., & Kalish, H. I. (1956). Discriminability and stimulus generalization. *Journal of Experimental Psychology, 51*(1), 79.
He, X., Yang, Y., Wang, L., & Yin, J. (2021). Tracking multiple perspectives: Spontaneous computation of what individuals in high entitative groups see. *Psychonomic Bulletin & Review, 28*, 879–887.
Henderson, J. M., Pollatsek, A., & Rayner, K. (1989). Covert visual attention and extrafoveal information use during object identification. *Perception & Psychophysics, 45*(3), 196–208.
Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (2001). The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences, 24*(5), 849–878.
Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*, 863.
Marquand, R. (1997). *Star wars: Return of the Jedi [film]*. Lucasfilm Ltd.
Müller, B. C., Oostendorp, A. K., Kühn, S., Brass, M., Dijksterhuis, A., & van Baaren, R. B. (2015). When triangles become human: Action co-representation for objects. *Interaction Studies, 16*(1), 54–67.
O'Grady, C., Scott-Phillips, T., Lavelle, S., & Smith, K. (2020). Perspective-taking is spontaneous but not automatic. *Quarterly Journal of Experimental Psychology, 73*(10), 1605–1628.
Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods, 162*(1–2), 8–13.
Phillips, B. (2021). Seeing seeing. *Philosophy and Phenomenological Research, 102*(1), 24–43.
Pick, D. F., Specker, S., Vu, K. P. L., & Proctor, R. W. (2014). Effects of face and inanimate-object contexts on stimulus–response compatibility. *Psychonomic Bulletin & Review, 21*, 376–383.
Rutkowska, N., Doradzińska, Ł., & Bola, M. (2022). Attentional prioritization of complex, naturalistic stimuli maintained in working-memory–A dot-probe event-related potentials study. *Frontiers in Human Neuroscience, 16*, Article 838338.
Salm-Hoogstraeten, S. V., & Müsseler, J. (2021). Human cognition in interaction with robots: Taking the robot's perspective into account. *Human Factors, 63*(8), 1396–1407.
Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & Bodley Scott, S. E. (2010). Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance, 36*(5), 1255.
Santiesteban, I., Catmur, C., Hopkins, S. C., Bird, G., & Heyes, C. (2014). Avatars and arrows: Implicit mentalizing or domain-general processing? *Journal of Experimental Psychology: Human Perception and Performance, 40*(3), 929.
Shepard, R. N. (1987). Towards a universal theory of generalization for psychological science. *Science, 237*(4820), 1317–1323.
Simon, J. R. (1969). Reactions toward the source of stimulation. *Journal of Experimental Psychology, 81*(1), 174.
Singh, A., Kalaichelvi, V., & Karthikeyan, R. (2022). A survey on vision guided robotic systems with intelligent control strategies for autonomous tasks. *Cogent Engineering, 9*(1), Article 2050020.
Surtees, A., Apperly, I., & Samson, D. (2016). I've got your number: Spontaneous perspective-taking in an interactive task. *Cognition, 150*, 43–52.
Tomasello, M. (2008). Origins of human cooperation. In *The Tanner lectures on human values* (pp. 77–80).
Tversky, B., & Hard, B. M. (2009). Embodied and disembodied cognition: Spatial perspective-taking. *Cognition, 110*(1), 124–129.
Van Breukelen, G. J. (2005). Psychometric modeling of response speed and accuracy with mixed and conditional regression. *Psychometrika, 70*, 359–376.
von Salm-Hoogstraeten, S., & Müsseler, J. (2021). Perspective taking while interacting with a self-controlled or independently-acting avatar. *Computers in Human Behavior, 118*, Article 106698.
von Salm-Hoogstraeten, S., Bolzius, K., & Müsseler, J. (2020). Seeing the world through the eyes of an avatar? Comparing perspective taking and referential coding. *Journal of Experimental Psychology: Human Perception and Performance, 46*(3), 264.
Wahn, B., & Berio, L. (2023). The influence of robot appearance on visual perspective taking: Testing the boundaries of the mere-appearance hypothesis. *Consciousness and Cognition, 116*, Article 103588.
Wahn, B., Berio, L., Weiß, M., & Newen, A. (2023). Try to see it my way: Humans take the level-1 visual perspective of humanoid robot avatars. *International Journal of Social Robotics*, 1–12.
Wegner, D. M., Schneider, D. J., Carter, S. R., & White, T. L. (1987). Paradoxical effects of thought suppression. *Journal of Personality and Social Psychology, 53*(1), 5.
Whelan, R. (2008). Effective analysis of reaction time data. *The Psychological Record, 58*, 475–482.
Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition, 69*(1), 1–34.
Wu, M. (2015). *Battle of balls (mobile version) [video game]*. Superpop&Lollipop.
Xiao, C., Xu, L., Sui, Y., & Zhou, R. (2021). Do people regard robots as human-like social partners? Evidence from perspective-taking in spatial descriptions. *Frontiers in Psychology, 11*, Article 578244.
Xiao, C., Fan, Y., Zhang, J., & Zhou, R. (2022). People do not automatically take the level-1 visual perspective of humanoid robot avatars. *International Journal of Social Robotics*, 1–12.
Yantis, S., & Egeth, H. E. (1999). On the distinction between visual salience and stimulus-driven attentional capture. *Journal of Experimental Psychology: Human Perception and Performance, 25*(3), 661–676.
Ye, T., Minato, T., Sakai, K., Sumioka, H., Hamilton, A., & Ishiguro, H. (2023). Human-like interactions prompt people to take a robot's perspective. *Frontiers in Psychology, 14*, Article 1190620.
Zhao, X., & Malle, B. F. (2022). Spontaneous perspective taking toward robots: The unique impact of humanlike appearance. *Cognition, 224*, Article 105076.
Zhao, M., Gersch, T. M., Schnitzer, B. S., Dosher, B. A., & Kowler, E. (2012). Eye movements and attention: The role of pre-saccadic shifts of attention in perception, memory and the control of saccades. *Vision Research, 74*, 40–60.
Zhao, X., Cusimano, C. J., & Malle, B. F. (2015, July). *In search of triggering conditions for spontaneous visual perspective taking*. CogSci.
Zhao, X., Cusimano, C., & Malle, B. F. (2015, March). Do people spontaneously take a robot's visual perspective?. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction extended abstracts* (pp. 133–134).
Zhao, X., Malle, B. F., & Gweon, H. (2016). *Is it a nine, or a six? Prosocial and selective perspective taking in four-year-olds*. CogSci.
Zhou, J., Peng, Y., Li, Y., Deng, X., & Chen, H. (2022). Spontaneous perspective taking of an invisible person. *Journal of Experimental Psychology: Human Perception and Performance, 48*(11), 1186.