International Journal of Human–Computer Interaction

ISSN: 1044-7318 (Print) 1532-7590 (Online) Journal homepage: www.tandfonline.com/journals/hihc20

How Can We Make Robots Intelligible? Building Perception-Behavior Causal Links for User-Centered Explainability

Xucong Hu, Enjie Xu, Haokui Xu, Mowei Shen & Jifan Zhou

To cite this article: Xucong Hu, Enjie Xu, Haokui Xu, Mowei Shen & Jifan Zhou (02 Mar 2026): How Can We Make Robots Intelligible? Building Perception-Behavior Causal Links for User-Centered Explainability, International Journal of Human–Computer Interaction, DOI: 10.1080/10447318.2026.2633220

To link to this article: https://doi.org/10.1080/10447318.2026.2633220

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2026.2633220

How Can We Make Robots Intelligible? Building Perception-Behavior Causal Links for User-Centered Explainability

Xucong Hu[a], Enjie Xu[a], Haokui Xu[b], Mowei Shen[a] and Jifan Zhou[a]

[a]Department of Psychology and Behavioral Sciences, Zhejiang University, Hangzhou, China; [b]Institute of Applied Psychology, College of Education, Zhejiang University of Technology, Hangzhou, China

ABSTRACT

As robots increasingly act as collaborative partners, their behavior must be not only functional but also understandable. Although robot control typically follows a Perception–Decision–Action (PDA) sequence, the causal link between perception and behavior is often not observable to users. To support explainability, users must intuitively grasp how perceptual input leads to action. Drawing on Hume's (1739) principles of causal perception—contiguity and contingency—we examined how temporal delays and perception–behavior alignment shape causal understanding. Results show that delays exceeding 600 ms, or actions preceding perception, disrupt perceived causality. Moreover, perception–behavior alignment must exceed 90% to maintain causal coherence and user trust. Together, these findings define empirically grounded design ranges: perceptual signals should precede actions by 0–600 ms, and alignment accuracy should remain above 90%. Meeting these criteria enhances robot explainability and improves human–robot interaction.

KEYWORDS

Human-robot interaction; causality; perspective-taking; user-centered explainability; mental model

1. Introduction

The rapid development of intelligent technologies is reshaping the field of human-robot interaction (HRI). As Goldberg (2019) notes, robotics has entered an era of collaborative intelligence, where robots are transitioning from isolated tools to integrated partners in human environments. This shift has changed expectations: robots are no longer just tools to perform tasks, but are now viewed as intelligent companions capable of human-like traits and behaviors. To achieve effective and seamless interaction with users, robot behavior must be explainable. Yet, this critical requirement has often been overlooked. Many robotic systems depend on opaque decision-making processes, such as end-to-end neural network models, which inherently lack transparency (Kim & Choi, 2021; Silva et al., 2023; Vadillo et al., 2024). Even when robot behaviors follow structured rules, these patterns are seldom communicated in ways that support user comprehension. Although humans are adept at inferring causal relationships from observed patterns, they are especially proficient at interpreting the behavior of other humans (Heberlein & Adolphs, 2005; Tomasello, 2008). The human mind is equipped with advanced social cognition mechanisms—including theory of mind, emotion recognition, and related processes—that allow for the effortless and automatic understanding of others' mental states (Baron-Cohen et al., 2013; Tomasello, 2008).

Previous studies have indicated several factors influence the human's perception of a robot's human-likeness. *Performance* is one consideration-robots that demonstrate human-level capabilities, such as D'Ambrosio et al. (2025) robot that won 45% of its table tennis matches against human opponents, can be perceived as more human-like (Aleksander, 2017; Bechar et al., 2009). *Appearance* also matters. Robots with more human-like physical features have been shown to inspire greater trust and more

CONTACT Mowei Shen mwshen@zju.edu.cn; Jifan Zhou jifanzhou@zju.edu.cn Department of Psychology and Behavioral Sciences, Zhejiang University, Zijingang Campus, 866 Yuhangtang Road, Hangzhou 310058, China; Haokui Xu haokuixu@zjut.edu.cn Institute of Applied Psychology, College of Education, Zhejiang University of Technology, 18 Chaowang Road, Hangzhou 310023, China

Supplemental data for this article can be accessed online at https://doi.org/10.1080/10447318.2026.2633220.

positive user experiences. For example, Van Pinxteren et al. (2019) found that human-like service robots led to higher user trust, a finding supported by Zhao and Malle (2022) and Kopp et al. (2022). The most critical factor, however, is the underlying information processing model that determines the pattern of robots' behavior. A *human-like cognitive framework*, such as Miller et al. (2017) perception-decision-action (PDA) chain, reinforces the impression of human-like responses. By following this model-perceiving the environment, processing data, and then acting-robots can be seen as social agents capable of interaction even if they don't appear human. This dynamic is highlighted in studies by von Salm-Hoogstraeten and Müsseler (2021) and Hu et al. (2025), which show that human-like behavioral patterns foster spontaneous social engagement, regardless of appearance or performance.

A fundamental question arises: what kind of information processing model can be considered comparable to human cognition? This question is intrinsically linked to explainability, a concept that has gained increasing significance in fields such as human-robot interaction, control engineering, and cognitive robotics (Jiang et al., 2022; Raees et al., 2024; Suffian et al., 2025). Explainability is typically defined as the degree to which a human observer can understand the reasoning behind an AI system's decision (Holzinger et al., 2019; Shneiderman, 2020). However, as the need for explainability varies across contexts, prior research distinguishes between two critical forms: Expert-Level Explainability and User-Level Explainability (Cabour et al., 2021; Liao & Varshney, 2021). Expert-Level Explainability is closely related to transparency, referring to whether a system's internal causal mechanisms and decision-making logic can be understood by engineers and AI specialists (Calisto et al., 2025; Gentile et al., 2025; Kim et al., 2023). However, for most non-expert users, User-Level Explainability—also referred to as the *mental model*—is of greater importance. It concerns whether users can intuitively understand and predict a system's behavior, even in the absence of technical knowledge about its underlying mechanisms (Liao & Varshney, 2021; Nacheva, 2015). Unlike expert-level transparency, which relies on analytical reasoning, user-level explainability emphasizes immediate perceptibility-allowing users to effortlessly grasp a robot's intentions without requiring inference or technical expertise. In our previous work, we proposed the *Social Agent Theory*, which demonstrated that robots are perceived as social entities—akin to humans—when they exhibit a specific set of features (Hu et al., 2025). These features include the capacity for both perception and behavior, corresponding to the input and output of information, respectively. Most critically, the presence of a clear and observable *causal* relationship between perception and behavior—where behavioral responses should be seen as a direct consequence of what the robot has sensed—is essential for supporting user-centered explainability. Given that robotic decision-making processes are often opaque, ensuring causal coherence enables robots to leverage humans' inherently powerful social cognitive mechanisms, such as perspective-taking, to promote intuitive understanding and interaction, thereby facilitating more natural collaboration.

Interestingly, the foundations of causal perception have been extensively studied in psychology and cognitive science. Hume (1739/1964) provided an early framework by identifying two core components of causality: contiguity and contingency. Contiguity refers to the temporal and spatial proximity of events-people intuitively perceive a causal relationship when one event closely follows another. For example, when flipping a light switch immediately turns on a light, a causal connection is naturally inferred. However, a long delay can weaken or eliminate this perception. Michotte's "launching effect" (1946/1963) demonstrated that when one moving object (A) hits another stationary object (B), causing B to move, people automatically perceive A as the cause. This automatic attribution highlights how certain event sequences inherently trigger causal perception, bypassing the need for deliberative reasoning. Contingency, on the other hand, relates to the consistency and reliability of the cause-effect relationship. If flipping a switch always turns on a light, the causal link is reinforced. Inconsistent outcomes, however, weaken the perceived connection. Research by Davis et al. (2020) confirmed that frequent and predictable co-occurrences between events strengthen causal perception. Together, contiguity and contingency form a framework for understanding how humans naturally perceive causality in their environment. Building on these principles, we propose that for an intelligent system to be perceived as causally reliable, its input-output interactions must satisfy both contiguity and contingency. The system's responses should be immediate and closely follow user inputs (contiguity) while remaining consistent and reliable over time (contingency). Meeting these conditions not only enhances user trust and

confidence in the system but also encourages the attribution of human-like qualities to the robot's behavior.

In human-robot interaction research, there are numerous approaches to determining whether users view a robot as a human-like collaborative partner. Traditionally, subjective measurement tools have been employed, including the Anthropomorphism Scale (Kamide et al., 2013; Spatola et al., 2021), the Collaborative Attitude Scale (Robert, 2021), and even the Turing Test (Turing, 2009). However, these methods are often influenced by individual biases, which can undermine their accuracy in reflecting users' actual cognitive states. An alternative approach is to observe behavioral responses to the robot's operational characteristics, which provides an objective and implicit measure. By focusing on users' actions, this approach minimizes subjective bias and offers a more direct window into their underlying cognitive processes. One particularly valuable phenomenon in this context is spontaneous perspective-taking (SPT). SPT occurs when users naturally adopt another entity's perspective, provided they perceive it as a social agent capable of interaction like humans. This process acts as a form of information preprocessing, paving the way for future collaboration (Hu et al., 2025; Zhou et al., 2022). SPT serves not only as a reliable indicator of whether users see a robot as a socially capable partner, but also as a widely applied measure in human-robot interaction studies (Hu & Tong, 2023; Salm-Hoogstraeten & Müsseler, 2021; Zhao et al., 2015; Zhao & Malle, 2022). For instance, Salm-Hoogstraeten and Müsseler (2021) used the avatar-Simon paradigm to determine whether users adopted a robot's perspective. Their findings showed that participants responded more quickly when the key press position matched the robot's location, reflecting SPT. This phenomenon offers a natural and implicit method to test whether individuals perceive a robot as social agent capable of interaction.

In this study, we also utilize the avatar-Simon paradigm (Salm-Hoogstraeten & Müsseler, 2021) to systematically examine the impact of time delays (contiguity) and consistent accuracy (contingency) between perception and behavior on causal perception and SPT effects in collaborative robots. Experiment 1 employs psychophysical methods to assess causal time contiguity by varying the delay between perception and action (−400 to 2000 ms), enabling us to identify how SPT effects change across different delay conditions. Experiment 2 investigates causal contingency by adjusting the frequency of the robot's correct behavior following perception (50% to 100%) and measuring the subsequent changes in SPT effects. By conducting these experiments, we aim to empirically characterize reference ranges that indicate the temporal and contingency conditions under which causal perception is most reliably observed. These findings will help refine the perception-decision-execution loop within parameters that are transparent and intuitive for users, ultimately enhancing the system's human-robot explainability.

2. Experiment 1

Experiment 1 adopted the avatar-Simon task procedure modified from Salm-Hoogstraeten and Müsseler (2021) to investigate how changes in the contiguity time delays between perception and behavior influence individuals' spontaneous perception of the robot as a social partner. The aim was to examine a range of time delays, identifying conditions where the effect emerges or disappears, thereby shedding light on when people are most likely to regard the robot as a social agent.

2.1. Method

2.1.1. Participants

Two hundred and sixteen participants (116 females; age: $M = 22.66$, $SD = 3.47$) were recruited on campus and compensated with 15 CNY or course credit for their participation. All participants had normal or corrected-to-normal vision and provided written informed consent for data collection, usage, and storage.

The required sample size was calculated using G*Power 3.1 (Faul et al., 2007). Based on effect sizes reported in prior studies (Böffel & Müsseler, 2019; Hu et al., 2025; Salm-Hoogstraeten & Müsseler, 2021), we performed power analyses with the following parameters: an effect size of 0.25, an alpha level of 0.05, a power of 0.98, and a mixed 9×2 design. The analyses indicated a minimum of 117

participants. However, to ensure comprehensive counterbalancing across multiple variables, a final sample size of 216 was selected, with 24 participants assigned to each between-subjects (time delay) condition. The study received approval from the Institutional Review Board at the Department of Psychology of the authors' university.

2.1.2. Apparatus and stimuli

The stimuli were created using PsychoPy (version 2023.2.3) (Peirce, 2007) and presented on a 16-inch monitor with a resolution of 1920 × 1080 pixels. Participants sat approximately 60 cm away from the screen and responded using the "q" and "p" keys on a standard keyboard, which were placed 8 cm on either side of the participant's midline and pressed with the left and right index fingers, respectively.

The target stimuli followed the design of Böffel and Müsseler (2019) and were adapted for the screen resolution used in this study. Each target appeared as a square, either dark blue (RGB 36, 115, 254) or light blue (RGB 98, 193, 254), measuring 77 pixels per side (approximately 1.34° visual angle). The targets were positioned 147 pixels (about 2.55°) above or below a central fixation cross, all displayed on an uniform gray background (RGB 155, 155, 155) spanning 1677 × 1258 pixels (approximately 29.26° × 21.94° visual angle).

The avatar used in this study—a virtual representation originally developed for VR and gaming environments (von Salm-Hoogstraeten & Müsseler, 2021)—was adapted from prior research (Hu et al., 2025). This robot representation has been extensively validated and has been shown to elicit spontaneous perspective-taking in users (Wahn et al., 2023; Wahn & Berio, 2023). The avatar was depicted as a black rectangle (364 × 264 pixels; approximately 4.61° × 7.08°) with two mechanical arms (264 × 406 pixels; 4.61° × 7.08°) and a camera head (301 × 227 pixels; 5.25° × 3.97°). To simulate perception, the camera head tilted 10 degrees toward the target, signaling its ability to perceive. To simulate behavior, one arm extended toward the target's center, stopping at a 49-pixel gap (0.86°), reflecting its capacity for performing a behavioral response (Figures 1 and 2).

2.2. Procedure and design

The experiment utilized a 9 (Time delay: −400, −200, 0, 200, 400, 600, 800, 1000, 2000 ms) × 2 (Compatibility: Compatible, Incompatible) mixed design. Time delay was the between-subjects variable, while compatibility served as the within-subjects variable. Trials were classified as *compatible* if the target's location matched the participant's response location from the avatar's perspective (e.g., when the avatar was positioned to the left (90°) and the participant pressed the left key for a light blue target, the target appeared on the avatar's left side). In contrast, trials were labeled as *incompatible* if the target's location from the avatar's perspective did not match the participant's response location (e.g., when the avatar was positioned to the left (90°) and the participant pressed the left key for a light blue target, the target appeared on the avatar's right side) (Figure 1). Better performance (e.g., faster reaction times) under compatible conditions than under incompatible conditions served as evidence that participants spontaneously adopted the avatar's perspective (Böffel & Müsseler, 2019; von Salm-Hoogstraeten & Müsseler, 2021). The measure of compatible advantage is commonly employed in such methods (Samson et al., 2010; Wahn et al., 2023; Zhao et al., 2015).

The experiment included two parts, each comprising 160 trials. On each trial, the target stimulus appeared randomly on either the left or right side of the avatar. Before starting each part, participants completed 20 practice trials, which were excluded from the main analysis. Participants' task was to classify the color of the target by pressing either the left or right button. They were informed that the robot executed other commands during the task that were unrelated to their own task. The initial positioning of the avatar (left or right) was counterbalanced across participants. The entire experiment lasted approximately 30 min.

Each trial began with a fixation cross and the avatar, both of which remained visible throughout the experiment. The fixation cross was displayed for 750 ms, after which the target appeared either above or below the fixation point. Participants were instructed to respond as quickly and accurately as possible. The timing of perception and behavior varied depending on the time delay condition. Positive time delays meant perception occurred first, lasting 200 to 2000 ms, followed by a 500 ms behavior

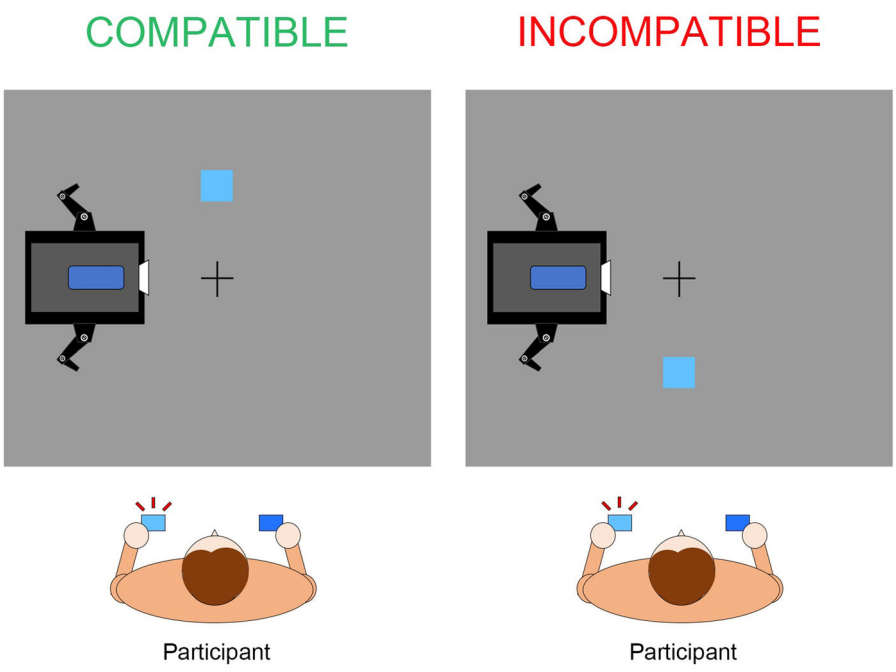

Figure 1. The schematic illustration of avatar-Simon task. *Note:* The figure illustrates the compatible (left) and incompatible (right) conditions in the avatar-Simon task (Böffel & Müsseler, 2018), using the avatar image from Hu et al. (2025). In the compatible condition, the target's position relative to the avatar matches the key's position relative to the participants; in the incompatible condition, they do not align.

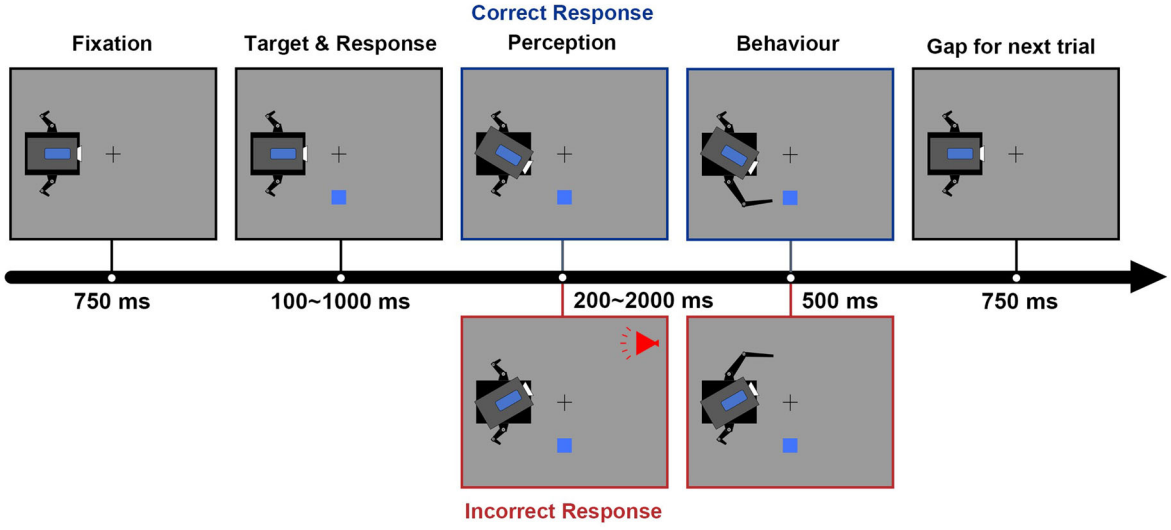

Figure 2. Experimental procedure of experiment 1. *Note:* A schematic illustration of a single trial in Experiment 1.

phase (Figure 2 illustrates this arrangement). Negative time delays reversed this sequence: behavior happened first for 200—400 ms, then perception followed for 500 ms (Figure 3). In the zero-delay condition, perception and behavior were simultaneous. Responses slower than 1,000 ms or faster than 100 ms were marked as errors, prompting the robot to provide incorrect feedback and emit an error tone. After each trial, the avatar returned to its initial state for 750 ms before the next trial began.

2.3. Result

False responses and outlier reaction times (RTs) outside the range of 100—1,000 ms were removed following established criteria (see Böffel & Müsseler, 2019). This exclusion resulted in the removal of

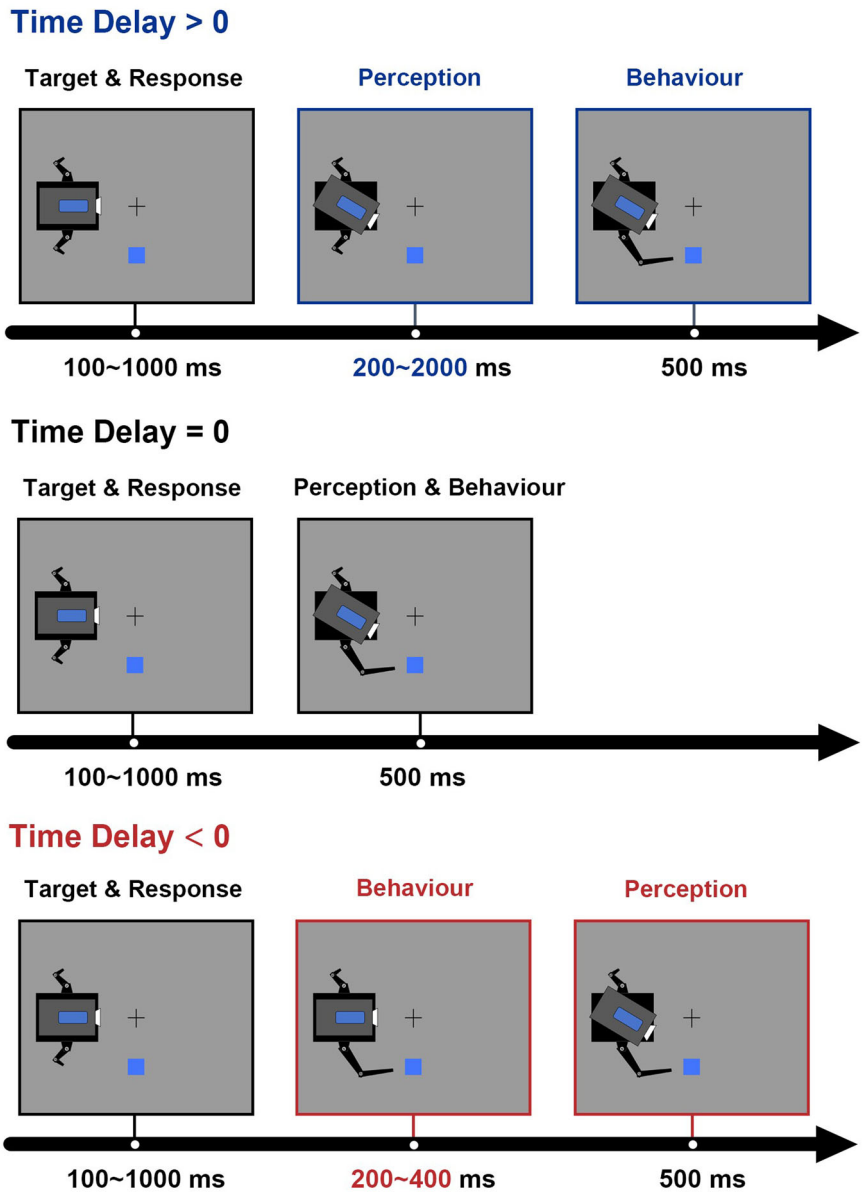

Figure 3. Experimental stimulus of experiment 1. *Note:* A schematic representation of the avatar's perception-behavior patterns under different time delays in Experiment 1.

3.72% of trials. A 9 (Time delay: −400, −200, 0, 200, 400, 600, 800, 1,000s, 2,000 ms) × 2 (Compatibility: Compatible, Incompatible) mixed ANOVA was performed, treating time delay as a between-subjects variable and compatibility as a within-subjects variable. The primary dependent variables-mean RTs and error rates-were analyzed separately. Bonferroni corrections were applied for *post hoc* comparisons.

To provide a clearer understanding of the effect sizes at different time delays, we calculated a "compatible advantage" measure, defined as the RT difference between compatible and incompatible conditions. A larger compatible advantage-indicated by faster RTs in compatible conditions-was interpreted as stronger evidence of spontaneous perspective-taking. This measure, serving as a descriptive statistic, has been used in prior research to quantify the extent of spontaneous perspective-taking (Böffel & Müsseler, 2019; von Salm-Hoogstraeten & Müsseler, 2021).

The analysis of variance (ANOVA) revealed a statistically significant main effect of compatibility ($F(1, 207) = 21.98$, $p < 0.001$, $\eta_p^2 = 0.096$), indicating that reaction times were faster under compatible conditions compared to incompatible conditions. Additionally, a significant main effect of time delay

was observed ($F(8, 207) = 2.77$, $p = 0.006$, $\eta_p^2 = 0.097$), suggesting that reaction times differed across the various time delays. Although the primary focus of the study was not the overall effect of time delay, detailed post hoc analyses are provided in the Supplementary Materials. A significant interaction between compatibility and time delay was also identified ($F(8, 207) = 2.37$, $p = 0.018$, $\eta_p^2 = 0.084$). Simple effects analyses indicated that at time delays of −400 ms and −200 ms, no significant differences were found between reaction times in compatible and incompatible conditions ($p = 0.676$ and $p = 0.359$, respectively). However, starting from a delay of 0 ms, reaction times under the compatible condition were significantly faster than those under the incompatible condition ($p < 0.001$), with a compatible advantage of 10 ms. Significant compatible advantages were also observed at 200 ms (9 ms, $p = 0.001$), 400 ms (7 ms, $p = 0.007$), and 600 ms (5 ms, $p = 0.047$). For longer delays-800 ms ($p = 0.764$), 1,000 ms ($p = 0.974$), and 2,000 ms ($p = 0.411$)-no significant differences in reaction times were detected (Figure 4).

For error rates (ER), we conducted a similar analysis to rule out speed-accuracy tradeoffs. The results showed no significant main effects of compatibility ($F(1, 207) = 1.67$, $p = 0.197$, $\eta_p^2 = 0.008$), time delay ($F(8, 207) = 1.37$, $p = 0.211$, $\eta_p^2 = 0.050$), or their interaction ($F(8, 207) = 1.49$, $p = 0.161$, $\eta_p^2 = 0.055$) (Figure 4). This pattern confirms that RT differences were not driven by changes in accuracy.

2.4. Discussion

Our results highlight that the temporal relationship between perception and behavior significantly impacts people's spontaneous perception of a robot as a social partner. Specifically, we found that spontaneous perspective-taking only occurred when the time delay was above zero, disappearing when the delay was negative. This suggests that the robot's perception must precede its behavior; if this sequence is disrupted, the robot's behavior no longer appears reasonable to participants. Moreover, we observed

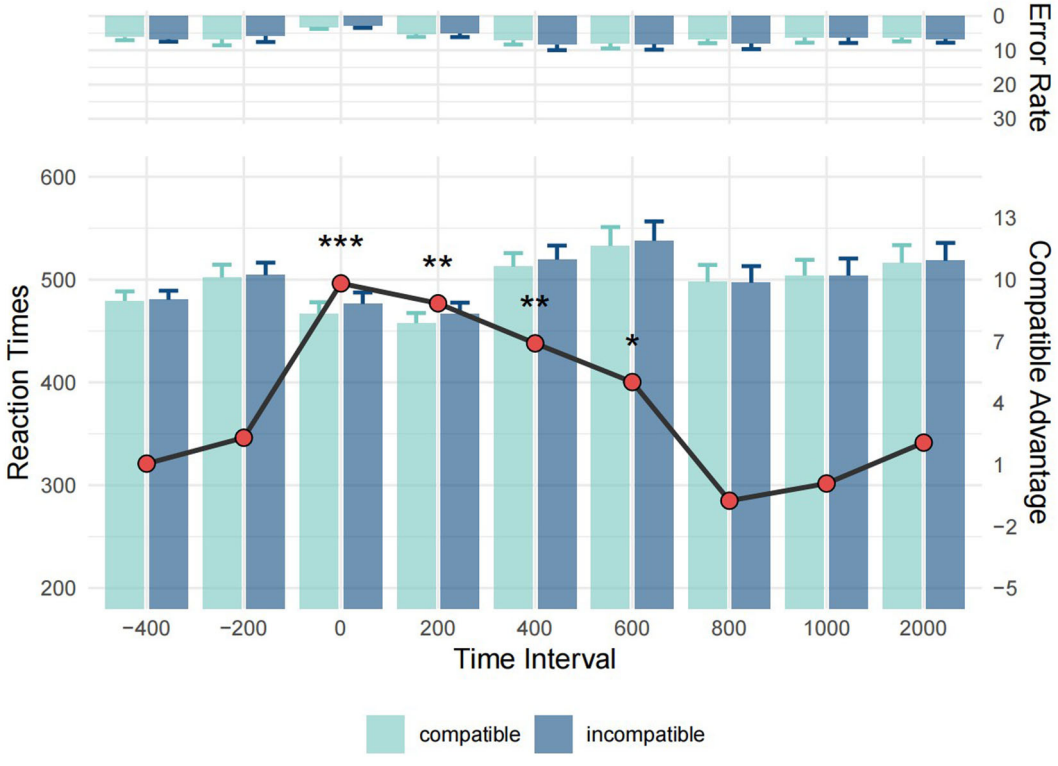

Figure 4. Results of experiment 1. *Note:* The figure displays the reaction times and error rates from Experiment 1 across the nine time delay conditions (−400, −200, 0, 200, 400, 600, 800, 1,000, and 2,000 ms). In panel (a), the bar plot shows mean reaction times for Compatible (light blue) and Incompatible (dark blue) conditions, with error bars representing standard errors of the mean. Overlaid on the bar plot is a black line with red circular markers, which indicates the compatible advantage-defined as the difference in reaction times between Incompatible and Compatible conditions-mapped onto the right y-axis. Asterisks denote statistically significant differences between conditions ($^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).

that the spontaneous perspective-taking effect persists within a time delay range of 0 to 600 ms, though it gradually diminishes as the delay increases. The compatible advantage decreased from 10 ms at a 0 ms delay to just 5 ms at 600 ms, indicating that the optimal perception-action interval falls within this window.

In the next experiment, we will explore how variations in contingency within an acceptable time delay range influence individuals' spontaneous perception of robots as social partners.

3. Experiment 2

Experiment 2 aimed to investigate how the frequency of correct behavior following perception-essentially, the *contingency* between perception and action-affects individuals' spontaneous perception of the robot as a social partner.

3.1. Method

3.1.1. Participants

Another 144 participants (62 females; age: $M = 23.30$, $SD = 4.16$) were recruited on campus and compensated with either 15 CNY or course credit. All participants had normal or corrected-to-normal vision and provided written informed consent. The required sample size was calculated using G*Power 3.1 (Faul et al., 2007) and effect sizes from previous studies (Böffel & Müsseler, 2019; Hu et al., 2025; Salm-Hoogstraeten & Müsseler, 2021), based on a 6 (Frequency: 100, 90, 80, 70, 60, 50%) × 2 (Compatibility: Compatible, Incompatible) design. The analysis indicated a minimum of 102 participants was necessary. To ensure thorough counterbalancing across multiple variables, a final sample size of 144 was selected, with 24 participants assigned to each frequency condition.

3.1.2. Apparatus and stimuli

The experimental setup largely followed that of Experiment 1, with adjustments only to the robot's perception-behavior feedback after the participant's response. In Experiment 2, the interval between the robot's perception and behavior was fixed at 200 ms. This duration, which exceeds the just noticeable difference, ensured a clear distinction between perception and behavior phases and was previously identified as a reasonable gap. In this experiment, perception always preceded behavior: the robot's perceptual process was displayed immediately after the participant's response, followed 200 ms later by its behavioral response. However, the correctness of these behavioral responses varied based on a frequency factor. For example, at 90% frequency, the robot performed the correct behavior 90% of the time. In the remaining 10%, the behavior was incorrect-such as reaching in a direction opposite to the target-making the behavior inconsistent with the perceived content (Figure 5). This manipulation was designed to reflect contingency-the frequency with which perception and correct behavior co-occurred.

3.2. Result

False responses and outlier reaction times (RTs) outside the range of 100—1,000 ms were removed following established criteria (see Böffel & Müsseler, 2019). This exclusion resulted in the removal of 6.83% of trials. A 6 (Frequency: 100, 90, 80, 70, 60, 50%) × 2 (Compatibility: Compatible, Incompatible) mixed ANOVA was performed, treating frequency as a between-subjects variable and compatibility as a within-subjects variable. The primary dependent variables-mean RTs and error rates-were analyzed separately. Bonferroni corrections were applied for post hoc comparisons.

The ANOVA revealed a statistically significant main effect of compatibility ($F(1, 138) = 10.19$, $p = 0.002$, $\eta_p^2 = 0.069$), indicating that RTs were faster under compatible conditions than under incompatible conditions. No significant main effect of frequency was observed ($F(5, 138) = 0.62$, $p = 0.686$, $\eta_p^2 = 0.022$). However, a significant interaction between compatibility and frequency emerged ($F(5, 138) = 2.64$, $p = 0.026$, $\eta_p^2 = 0.087$). Simple effects analyses showed that at a frequency of 100%, RTs were significantly faster under the compatible condition than under the incompatible condition, with a 10 ms advantage ($p < 0.001$). A similar significant difference was found at 90% frequency, with a 7 ms

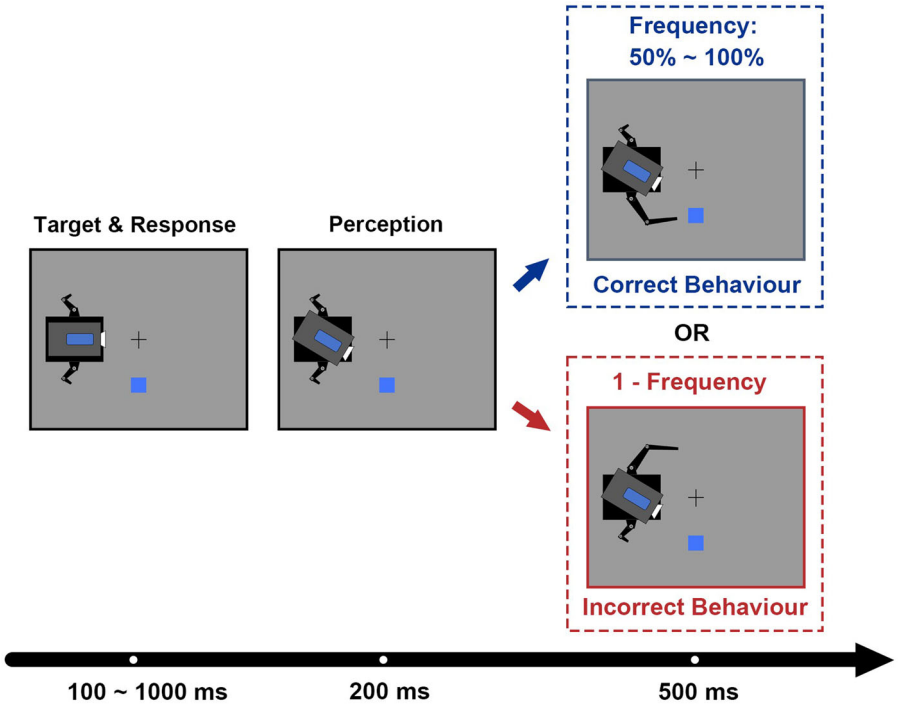

Figure 5. Experimental stimulus and procedure of Experiment 2. *Note:* A schematic representation of the avatar's perception-behavior patterns under different accuracy in Experiment 2.

advantage ($p = 0.012$). When the frequency of correct behavior decreased further, no significant compatibility advantage was observed at 80% ($p = 0.802$), 70% ($p = 0.793$), 60% ($p = 0.503$), or 50% ($p = 0.793$) (Figure 6).

The analysis of error rates (ER) also revealed no significant main effects of compatibility ($F(1, 138) = 2.16$, $p = 0.144$, $\eta_p^2 = 0.015$), frequency ($F(5, 138) = 0.48$, $p = 0.792$, $\eta_p^2 = 0.017$), or their interaction ($F(5, 138) = 0.83$, $p = 0.528$, $\eta_p^2 = 0.029$) (Figure 6). This pattern confirms that RT differences were not driven by changes in accuracy.

3.3. Discussion

Our findings demonstrate that contingency-the frequency of correct behaviors following perception-plays a crucial role in whether individuals spontaneously perceive a robot as a social partner. When the robot's corresponding behaviors occurred 90 or 100% of the time after perception, participants consistently adopted the robot's perspective. However, as this frequency decreased, the effect dissipated. These results suggest that a robot's behavioral accuracy of at least 90% following perception is necessary to reliably trigger people's spontaneous perspective-taking.

4. General discussion

The current study examined how the temporal contiguity (time delays) and contingency (frequency of correct behavior) between a robot's perception and its behavior influence humans' spontaneous consideration of the robot as a social partner. The results showed that people spontaneously adopted the robot's perspective when the time delay between perception and behavior ranged from 0 to 600 ms. However, if behavior preceded perception or the delay exceeded this range, the perspective-taking effect disappeared. Additionally, the effect was sustained only when the frequency of correct behaviors following perception reached at least 90%. If the frequency dropped below 90%, the effect vanished again. These findings delineate empirically grounded design reference ranges for robotic engineering and interpretable goal-driven AI, showing that causal interpretation is most reliable when robots' perception

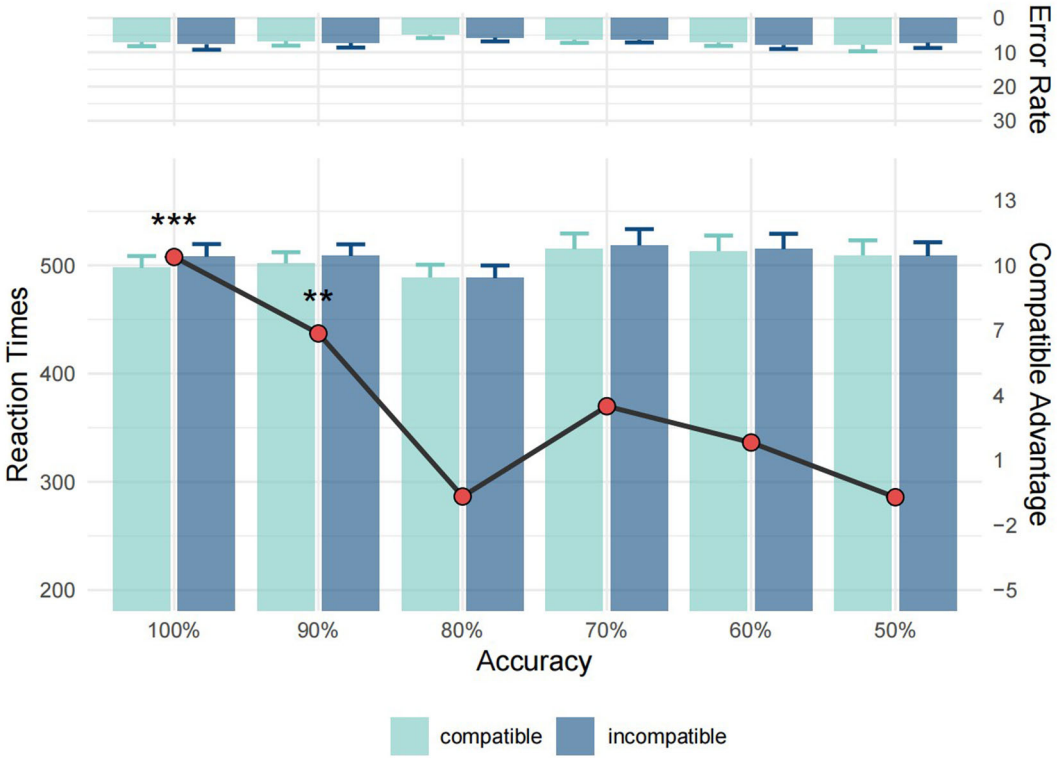

Figure 6. Results of experiment 2. *Note:* The figure displays the reaction times and error rates from Experiment 2, separated by the six frequency conditions (100, 90, 80, 70, 60, 50%) and by compatibility (light blue for Compatible, dark blue for Incompatible conditions). The plotting conventions align with those presented in Figure 4.

precedes behavior by approximately 0–600 ms and when perception–behavior alignment remains high (above 90%).

However, certain issues from previous research warrant further discussion. For instance, in Experiment 1, participants spontaneously adopted the robot's perspective even when perception and behavior occurred simultaneously (0 ms delay). At first glance, this observation may seem to challenge the conventional assumption that perception must precede behavior in a causal sequence. However, this finding is consistent with earlier work (Hu et al., 2025), which highlights that while perception typically precedes behavior, the underlying perceptual processes driving behavior do not always result in immediately observable changes. In many cases, the underlying perceptual processing remains covert, even when the external events appear simultaneous. This perspective aligns with prior research demonstrating that subtle pre-attentive shifts or preparatory processes often occur prior to overt perception and behavior, even if they are not always detected (Deubel & Schneider, 1996; Henderson et al., 1989; Zhao et al., 2012). This preprocessing stage enables rapid attentional shifts and response initiation, particularly under conditions that demand swift reactions. Externally, these events can appear nearly simultaneous-similar to a baseball player glancing at a ball and almost instantly beginning their swing.

An alternative explanation of the SPT effect observed in the current study is that it may be driven by simple action-based causality alone—such that reaching behavior, if causally contingent, would be sufficient to elicit SPT, without requiring an explicit perception–behavior sequence. To examine this possibility, we conducted a supplementary control experiment. Specifically, we tested whether SPT could emerge when the robot retained a perceptual structure (a camera) but did not execute any perceptual movement (e.g., camera movement), while preserving the strongest possible causal parameters (0 ms delay and 100% contingency). Under these conditions, the SPT effect was absent (see Supplementary Materials for details). This finding provides further evidence that SPT critically depends on an observable perception–behavior sequence, rather than on simple action-based causality alone.

A key question emerging from these findings is why the observed time delay and frequency thresholds align with specific ranges. Previous research has extensively examined the temporal constraints of

causal perception. For instance, studies on the "launching effect" demonstrate that a causal relationship between two events-such as ball A striking ball B and ball B moving-is only perceived if the events occur within a very short interval, typically less than 100 ms (Bechlivanidis & Lagnado, 2016; Cravo et al., 2015; Michotte, 2017). However, the temporal relationship between an agent's perception and its subsequent behavioral response likely follows a different model. Here, delay is not merely the temporal gap between two physical events; rather, it represents the interval between perceiving information, forming an intention, and executing a response. This sequence involves multiple cognitive stages and is inherently more complex and prolonged than the simple mechanical interaction between two objects. For example, Libet et al. (1993) found that the time needed for a subjective intention to initiate movement is typically around 200–350 ms. In our experiments, participants interpreted the agent's perception and subsequent behavior as causally connected, perceiving the agent as a potential social partner. Notably, our findings suggest that humans exhibit greater tolerance for causal intervals when interacting with social agents than what is typically observed in traditional causal perception studies. This broader tolerance may stem from humans' long evolutionary history of interactions with non-human social partners, such as domesticated animals, which often exhibit delayed responses-for example, pigeons taking a brief moment before reacting (Blough, 2000). This flexibility may reflect humans' capacity for learned adaptation, as prior research suggests that exposure to slower agents can implicitly adjust people's own temporal expectations (Bargh et al., 1996). This evolutionary and experiential context could explain why humans are more accepting of extended causal intervals when evaluating social agents. Importantly, these evolutionary and experiential considerations are offered only as post hoc interpretations of the empirical results obtained via the fixed-window psychophysical approach, rather than as motivations for the experimental design or the time window itself. Identifying the precise underlying mechanisms remains an open question for future research.

Secondly, regarding frequency, our findings reveal that participants required a remarkably high level of accuracy—around 90%—to perceive a causal link between the agent's perception and its subsequent behavior. This threshold is notably higher than what has been reported in many classic studies on causal perception. For example, previous research often finds that individuals can detect causal relationships at accuracy levels around 50–60% (Hommel et al., 2003; Msetfi et al., 2013). Such studies typically involve simple action-effect tasks, such as pressing a button and observing whether a light turns on. By contrast, evaluating an agent's behavior may be more sensitive. When people judge whether an agent is reliably responding to perceived information, the stakes are arguably higher: misjudging the agent's competence could lead to social or safety risks, especially in collaborative or interactive settings, where even small errors in machine decision-making can significantly undermine user trust (Fan et al., 2008; Kim et al., 2025). This heightened sensitivity may explain why people adopt a stricter criterion for detecting causal coherence in agent behavior compared to simpler tasks. Future research should clarify the cognitive mechanisms underlying these thresholds and determine how anthropomorphism, context, and risk perceptions influence people's evaluations of social agents.

Notably, although the compatibility advantage appeared to increase slightly at the 70% accuracy condition, post hoc comparisons with neighboring conditions (50, 60, and 80%) revealed no statistically significant differences (see Supplementary Materials for details). We therefore interpret this apparent increase as a nonsystematic fluctuation rather than a reliable effect. When information is uncertain and the perception-behavior relationship is unstable, people may occasionally over-attribute causal structure—a phenomenon related to the illusion of causality (Alloy & Abramson, 1979). At present, our data do not allow us to determine why such a fluctuation would emerge at a specific level (e.g., 70% rather than 60 or 80%). Identifying the mechanisms underlying this pattern remains an open question for future research.

Moreover, it is important to note that the temporal and contingency thresholds identified in the present study were derived from a relatively homogeneous sample of young university adults, a population that typically exhibits higher familiarity with digital technologies and robotic systems. Prior research indicates that age and technology experience systematically shape human-robot interaction. Compared with younger adults, older adults tend to report lower acceptance of robots and show greater sensitivity to behavioral inconsistency, particularly when system behavior is difficult to interpret (Broadbent et al.,

2009; Czaja et al., 2006; Scopelliti et al., 2005). In contrast, children may display the opposite pattern, showing greater tolerance toward robot errors and behavioral inconsistency (Breazeal et al., 2016). Moreover, technology experience has been shown to modulate causal and intentional interpretations of robot behavior: limited prior experience is associated with stricter judgment criteria, whereas greater familiarity promotes more tolerant and flexible interpretations (Heerink et al., 2010). Accordingly, the reference ranges reported here should be interpreted as estimates applicable to young adults with relatively high ease of interaction with technology and robotics. Future research should examine whether these thresholds shift across age groups, levels of robot exposure, and individual differences in technology attitudes.

The main contribution of this study is the empirical delineation of reference ranges for the temporal and contingency conditions under which robot behavior is interpreted as causal by human observers. Building upon our previous work (Hu et al., 2025), we emphasize the importance of integrating an overt perception module as a standard feature of intelligent systems. Even in robots that do not rely on directional perception—such as those equipped with omnidirectional LiDAR—a virtual sensory indicator can substantially improve interaction quality. Importantly, this module does not need to perform actual sensing functions; rather, it should provide a clear, perceivable cue to the user, such as light signals or symbolic eyes that emulate human-like sensory behaviors. For instance, Lasota et al. (2017) reviewed human-robot collaboration in manufacturing and highlighted that workers often cannot predict robot motions, especially when robots lack expressive cues, leading to reduced safety and efficiency. In addition to the presence of perceptual cues, it is critical that the robot maintains a salient focus on one target when interacting with multiple objects or individuals. This selective focus allows users to infer the robot's attentional priorities and reduces ambiguity during interaction. The importance of this principle is evident in real-world service environments; for example, hotel service robots frequently issue generic commands like "please make way" without orienting toward a specific person, leaving guests confused about who is being addressed. Furthermore, the temporal relationship between perception and action must be reliably perceivable. The perceptual cue should consistently shift toward the intended target approximately 0–200 milliseconds before the initiation of a behavior. While this temporal delay does not need to reflect real sensory processing, it should reliably occur before the action to support causal inference. Critically, maintaining at least 90% consistency between perceptual cues and subsequent actions is necessary to sustain user trust. This requirement becomes particularly essential in high-stakes decision-making environments, such as autonomous driving, where users must be able to anticipate system behavior to maintain situational awareness and ensure safe collaboration (Wiegand et al., 2019). Our approach has already been validated in ecological, real-world scenarios (Salm-Hoogstraeten & Müsseler, 2021), and we encourage future research to replicate and expand upon these findings under naturalistic conditions. This not only improves usability and user trust but also lays the groundwork for seamless human-robot collaboration.

Author contributions

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

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was supported by the National Natural Science Foundation of China [Grants 32371088, 62337001, and 32471093], the Zhejiang Provincial Natural Science Foundation of China [Grant No. LQN26C090003], Space Medical Experiment Project of CMSP [HYZHXMR01003], and the Fundamental Research Funds for the Central Universities [226-2024-00118].

ORCID

Xucong Hu http://orcid.org/0009-0001-0711-1577
Enjie Xu http://orcid.org/0009-0007-1750-2137
Haokui Xu http://orcid.org/0000-0003-4259-134X
Mowei Shen http://orcid.org/0000-0001-7661-2968
Jifan Zhou http://orcid.org/0000-0003-0166-7125

Data availability statement

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

References

Aleksander, I. (2017). Partners of humans: A realistic assessment of the role of robots in the foreseeable future. *Journal of Information Technology*, *32*(1), 1–9. https://doi.org/10.1057/s41265-016-0032-4

Alloy, L. B., & Abramson, L. Y. (1979). Judgment of contingency in depressed and nondepressed students: Sadder but wiser? *Journal of Experimental Psychology. General*, *108*(4), 441–485. https://doi.org/10.1037/0096-3445.108.4.441

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*(2), 230–244. https://doi.org/10.1037/0022-3514.71.2.230

Baron-Cohen, S., Tager-Flusberg, H., & Lombardo, M. (Eds.) (2013). *Understanding other minds: Perspectives from developmental social neuroscience*. OUP Oxford.

Bechar, A., Meyer, J., & Edan, Y. (2009). An objective function to evaluate performance of human–robot collaboration in target recognition tasks. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, *39*(6), 611–620. https://doi.org/10.1109/tsmcc.2009.2020174

Bechlivanidis, C., & Lagnado, D. A. (2016). Time reordered: Causal perception guides the interpretation of temporal order. *Cognition*, *146*, 58–66. https://doi.org/10.1016/j.cognition.2015.09.001

Blough, D. S. (2000). Effects of priming, discriminability, and reinforcement on reaction-time components of pigeon visual search. *Journal of Experimental Psychology. Animal Behavior Processes*, *26*(1), 50–63. https://doi.org/10.1037/0097-7403.26.1.50

Böffel, C., & Müsseler, J. (2018). Perceived ownership of avatars influences visual perspective taking. *Frontiers in Psychology*, *9*, 743. https://doi.org/10.3389/fpsyg.2018.00743

Böffel, C., & Müsseler, J. (2019). Visual perspective taking for avatars in a Simon task. *Attention, Perception & Psychophysics*, *81*(1), 158–172. https://doi.org/10.3758/s13414-018-1573-0

Breazeal, C., Harris, P. L., DeSteno, D., Kory Westlund, J. M., Dickens, L., & Jeong, S. (2016). Young children treat robots as informants. *Topics in Cognitive Science*, *8*(2), 481–491. https://doi.org/10.1111/tops.12192

Broadbent, E., Stafford, R., & MacDonald, B. (2009). Acceptance of healthcare robots for the older population: Review and future directions. *International Journal of Social Robotics*, *1*(4), 319–330. https://doi.org/10.1007/s12369-009-0030-6

Cabour, G., Morales, A., Ledoux, É., & Bassetto, S. (2021). Towards an explanation space to align humans and explainable-AI teamwork. *arXiv preprint arXiv:2106.01503*. https://arxiv.org/abs/2106.01503

Calisto, F. M., Abrantes, J. M., Santiago, C., Nunes, N. J., & Nascimento, J. C. (2025). Personalized explanations for clinician-AI interaction in breast imaging diagnosis by adapting communication to expertise levels. *International Journal of Human-Computer Studies*, *197*, 103444. https://doi.org/10.1016/j.ijhcs.2025.103444

Cravo, A. M., Santos, K. M., Reyes, M. B., Caetano, M. S., & Claessens, P. M. (2015). Visual causality judgments correlate with the phase of alpha oscillations. *Journal of Cognitive Neuroscience*, *27*(10), 1887–1894. https://doi.org/10.1162/jocn_a_00832

Czaja, S. J., Charness, N., Fisk, A. D., Hertzog, C., Nair, S. N., Rogers, W. A., & Sharit, J. (2006). Factors predicting the use of technology: Findings from the Center for Research and Education on Aging and Technology Enhancement (CREATE). *Psychology and Aging*, *21*(2), 333–352. https://doi.org/10.1037/0882-7974.21.2.333

D'Ambrosio, D. B., Abeyruwan, S. W., Graesser, L., Iscen, A., Amor, H. B., Bewley, A., Reed, B. J., Reymann, K., Takayama, L., Tassa, Y., Choromanski, K., Coumans, E., Jain, D., Jaitly, N., Jaques, N., Kataoka, S., Kuang, Y., Lazic, N., Mahjourian, R., … Sanketi, P. R. (2025, May). Achieving human level competitive robot table tennis. In *Proceedings of the 2025 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 74–82). IEEE. https://doi.org/10.1109/ICRA55743.2025.11127501

Davis, Z. J., Bramley, N. R., & Rehder, B. (2020). Causal structure learning in continuous systems. *Frontiers in Psychology*, *11*, 244. https://doi.org/10.3389/fpsyg.2020.00244

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research*, *36*(12), 1827–1837. https://doi.org/10.1016/0042-6989(95)00294-4

Fan, X., Oh, S., McNeese, M., Yen, J., Cuevas, H., Strater, L., & Endsley, M. R. (2008, January). The influence of agent reliability on trust in human-agent collaboration. In *Proceedings of the 15th European Conference on Cognitive Ergonomics: The Ergonomics of Cool Interaction (ECCE 2008)* (pp. 1–8). Association for Computing Machinery. https://doi.org/10.1145/1473018.1473028

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. https://doi.org/10.3758/BF03193146

Gentile, D., Donmez, B., & Jamieson, G. A. (2025). Human performance effects of combining counterfactual explanations with normative and contrastive explanations in supervised machine learning for automated decision assistance. *International Journal of Human-Computer Studies*, *196*, 103434. https://doi.org/10.1016/j.ijhcs.2024.103434

Goldberg, K. (2019). Robots and the return to collaborative intelligence. *Nature Machine Intelligence*, *1*(1), 2–4. https://doi.org/10.1038/s42256-018-0008-x

Heberlein, A. S., & Adolphs, R. (2004). Functional anatomy of human social cognition. In A. Easton & N. J. Emery (Eds.), *The Cognitive Neuroscience of Social Behaviour* (pp. 169–206), Psychology Press. https://doi.org/10.4324/9780203311875-13

Heerink, M., Kröse, B., Evers, V., & Wielinga, B. (2010). Assessing acceptance of assistive social agent technology by older adults: The Almere model. *International Journal of Social Robotics*, *2*, 361–375. https://doi.org/10.1007/s12369-010-0068-5

Henderson, J. M., Pollatsek, A., & Rayner, K. (1989). Covert visual attention and extrafoveal information use during object identification. *Perception & Psychophysics*, *45*(3), 196–208. https://doi.org/10.3758/bf03210697

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery*, *9*(4), e1312. https://doi.org/10.1002/widm.1312

Hommel, B. (2003). Event files: Evidence for automatic integration of stimulus–response features. *Psychological Research*, *67*, 89–103. https://doi.org/10.1007/s00426-002-0151-3

Hu, X., & Tong, S. (2023). Effects of robot animacy and emotional expressions on perspective-taking abilities: A comparative study across age groups. *Behavioral Sciences*, *13*(9), 728. https://doi.org/10.3390/bs13090728

Hu, X., Xu, H., Chen, H., Shen, M., & Zhou, J. (2025). Good to see you R2-D2: Inducing spontaneous perspective-taking towards non-human agents through human-like gaze and reach. *Cognition*, *259*, 106101. https://doi.org/10.1016/j.cognition.2025.106101

Hume, D. (1739/1964). *A treatise of human nature*. Oxford University Press.

Jiang, J., Kahai, S., & Yang, M. (2022). Who needs explanation and when? Juggling explainable AI and user epistemic uncertainty. *International Journal of Human-Computer Studies*, *165*, 102839. https://doi.org/10.1016/j.ijhcs.2022.102839

Kamide, H., Eyssel, F., & Arai, T. (2013). Psychological anthropomorphism of robots. In G. Herrmann, M. J. Pearson, A. Lenz, P. Bremner, A. Spiers, & U. Leonards (Eds.), *Social Robotics: ICSR 2013* (Lecture Notes in Computer Science, Vol. 8239, pp. 199–208). Springer International Publishing. https://doi.org/10.1007/978-3-319-02675-6_20

Kim, J. Y., Lester, C., & Yang, X. J. (2025). Beyond binary decisions: Evaluating the effects of AI error type on trust and performance in AI-assisted tasks. *human Factors*, *67*(10), 1062–1083. https://doi.org/10.1177/00187208251326795

Kim, S., & Choi, J. (2021, September). Explaining the decisions of deep policy networks for robotic manipulations. In *Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 2663–2669). IEEE. https://doi.org/10.1109/IROS51168.2021.9636594

Kim, S., Choo, S., Park, D., Park, H., Nam, C. S., Jung, J. Y., & Lee, S. (2023). Designing an XAI interface for BCI experts: A contextual design for pragmatic explanation interface based on domain knowledge in a specific context. *International Journal of Human-Computer Studies*, *174*, 103009. https://doi.org/10.1016/j.ijhcs.2023.103009

Kopp, T., Baumgartner, M., & Kinkel, S. (2022). How linguistic framing affects factory workers' initial trust in collaborative robots: The interplay between anthropomorphism and technological replacement. *International Journal of Human-Computer Studies*, *158*, 102730. https://doi.org/10.1016/j.ijhcs.2021.102730

Lasota, P. A., Fong, T., & Shah, J. A. (2017). A survey of methods for safe human-robot interaction. *Foundations and Trends® in Robotics*, *5*(4), 261–349. https://doi.org/10.1561/2300000052

Liao, Q. V., & Varshney, K. R. (2021). Human-centered explainable AI (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790*. https://arxiv.org/abs/2110.10790

Libet, B., Gleason, C. A., Wright, E. W., & Pearl, D. K. (1993). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). In R. Miller (Ed.), *Neurophysiology of consciousness.* pp. 249–268). Birkhäuser Boston.

Michotte, A. (2017). *The perception of causality*. Routledge.

Miller, G. A., Eugene, G., & Pribram, K. H. (2017). Plans and the structure of behaviour. In W. Buckley (Ed.), *Systems research for behavioral science: A sourcebook* (pp. 369–382). Routledge.

Msetfi, R. M., Wade, C., & Murphy, R. A. (2013). Context and time in causal learning: Contingency and mood dependent effects. *PLOS One*, *8*(5), e64063. https://doi.org/10.1371/journal.pone.0064063

Nacheva, R. (2015). The importance of users' mental models for developing usable human-machine interfaces. *Scientific Papers at the University of Rousse*, *54*(6.1), 132–135. https://doi.org/10.13140/RG.2.1.4329.3844

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1–2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Raees, M., Meijerink, I., Lykourentzou, I., Khan, V. J., & Papangelis, K. (2024). From explainable to interactive AI: A literature review on current trends in human-AI interaction. *International Journal of Human-Computer Studies*, *189*, 103301. https://doi.org/10.1016/j.ijhcs.2024.103301

Robert, L. P. Jr. (2021, July). *A measurement of attitude toward working with robots (awro): A compare and contrast study of AWRO with negative attitude toward robots (nars)* [Paper presentation]. International Conference on Human-Computer Interaction, Cham (pp. 288–299). Springer International Publishing.

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. https://doi.org/10.1177/0018720820933764

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), 1255–1266. https://doi.org/10.1037/a0018729

Scopelliti, M., Giuliani, M. V., & Fornara, F. (2005). Robots in a domestic setting: A psychological approach. *Universal Access in the Information Society*, *4*(2), 146–155. https://doi.org/10.1007/s10209-005-0118-1

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

Silva, A., Schrum, M., Hedlund-Botti, E., Gopalan, N., & Gombolay, M. (2023). Explainable artificial intelligence: Evaluating the objective and subjective impacts of xai on human-agent interaction. *International Journal of Human–Computer Interaction*, *39*(7), 1390–1404. https://doi.org/10.1080/10447318.2022.2101698

Spatola, N., Kühnlenz, B., & Cheng, G. (2021). Perception and evaluation in human–robot interaction: The Human–Robot Interaction Evaluation Scale (HRIES)—a multicomponent approach of anthropomorphism. *International Journal of Social Robotics*, *13*(7), 1517–1539. https://doi.org/10.1007/s12369-020-00667-4

Suffian, M., Kuhl, U., Bogliolo, A., & Alonso-Moral, J. M. (2025). The role of user feedback in enhancing understanding and trust in counterfactual explanations for explainable AI. *International Journal of Human-Computer Studies*, *199*, 103484. https://doi.org/10.1016/j.ijhcs.2025.103484

Tomasello, M. (2008). Origins of human cooperation. In S. M. McMurrin (Ed.), *The Tanner lectures on human values* (pp. 77–80). Cambridge University Press.

Turing, A. M. (2009). *Computing machinery and intelligence* (pp. 23–65). Springer Netherlands.

Vadillo, J., Santana, R., Lozano, J. A., & Kwiatkowska, M. (2024). Uncertainty-aware explanations through probabilistic self-explainable neural networks. *arXiv preprint arXiv:2403.13740*. https://arxiv.org/abs/2403.13740

Van Pinxteren, M. M., Wetzels, R. W., Rüger, J., Pluymaekers, M., & Wetzels, M. (2019). Trust in humanoid robots: Implications for services marketing. *Journal of Services Marketing*, *33*(4), 507–518. https://doi.org/10.1108/jsm-01-2018-0045

von Salm-Hoogstraeten, S., & Müsseler, J. (2021). Perspective taking while interacting with a self-controlled or independently-acting avatar. *Computers in Human Behavior*, *118*, 106698. https://doi.org/10.1016/j.chb.2021.106698

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*, 103588. https://doi.org/10.1016/j.concog.2023.103588

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*, *17*, 523–534. https://doi.org/10.1007/s12369-023-01036-7

Wiegand, G., Schmidmaier, M., Weber, T., Liu, Y., & Hussmann, H. (2019, May). I drive-you trust: Explaining driving behavior of autonomous cars. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)* (pp. 1–6). Association for Computing Machinery. https://doi.org/10.1145/3290607.3312817

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. https://doi.org/10.1016/j.visres.2012.06.017

Zhao, X., & Malle, B. F. (2022). Spontaneous perspective taking toward robots: The unique impact of humanlike appearance. *Cognition*, *224*, 105076. https://doi.org/10.1016/j.cognition.2022.105076

Zhao, X., Cusimano, C., & Malle, B. F. (2015). In search of triggering conditions for spontaneous visual perspective taking. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (vol. 37).

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), 1186–1200. https://doi.org/10.1037/xhp0001047

About the authors

Xucong Hu is currently a Master's student in Psychology at Zhejiang University. His research focuses on social cognition, visual perspective-taking, and human-robot interaction, with an emphasis on perception-behavior coupling and explainable social behavior in artificial agents.

Enjie Xu is currently a PhD student in Psychology at Zhejiang University. His research focuses on cognitive modeling and visual cognition, with particular interest in animacy perception, agency attribution, and their underlying computational mechanisms.

Haokui Xu is an Assistant Professor at the Institute of Applied Psychology, College of Education, Zhejiang University of Technology. He received his PhD in Psychology from Zhejiang University. His research focuses on visual cognition, social cognition, and computational modeling of cognitive processes.

Mowei Shen is a Professor in the Department of Psychology and Behavioral Sciences at Zhejiang University. He received his PhD in Psychology from Hangzhou University. His research focuses on cognitive processes and modeling, social cognition, and human factors engineering.

Jifan Zhou is a Professor in the Department of Psychology and Behavioral Sciences at Zhejiang University. He received his PhD in Psychology from Zhejiang University. His research focuses on social cognition, working memory, and computational approaches to human cognition.