Do People Spontaneously Take a Robot's Visual Perspective?

Xuan Zhao
Department of Cognitive, Linguistic, and Psychological Sciences, Brown University
Providence, RI, USA
Email: xuan_zhao@brown.edu

Corey Cusimano
Department of Psychology
University of Pennsylvania
Philadelphia, PA, USA

Bertram F. Malle
Department of Cognitive, Linguistic, and Psychological Sciences, Brown University
Providence, RI, USA
Email: bfmalle@brown.edu

***Abstract*— Visual perspective taking plays a fundamental role in both human-human interaction and human-robot interaction (HRI). In three experiments, we took a novel approach to the topic of visual perspective taking in HRI, examining whether, and under what conditions, people spontaneously take a robot's visual perspective. Using two different robot models, we found that specific behaviors performed by a robot—namely, object-directed gaze and goal-directed reaching—led many human viewers to take the robot's visual perspective, though slightly fewer than when the same behaviors were performed by a person. However, we found no difference in people's perspective-taking tendency toward robots that differed in their human-likeness. Also, reaching became an especially effective perspective-taking trigger when it was displayed in a video rather than in a photograph. Taken together, these findings suggest that certain nonverbal behaviors in robots are sufficient to trigger the mechanism of mental state attribution—visual perspective taking in particular—in human observers. Therefore, people's spontaneous perspective-taking tendencies should be taken into account when designing intuitive and effective human-centered robots.**

Keywords*— Human-robot interaction (HRI); communication; perspective taking; nonverbal behaviors; mind perception; humanoid robot.

I. INTRODUCTION

Whereas Alan Turing's famous question, "Can machines think?" [1] is not easily answerable by empirical methods, the question of whether people think that machines can think is tractable by empirical research. By investigating people's assumptions and perceptions of robots' mental capacities, such research helps to elucidate how humans perceive other minds; predicts how people will naturally interact with robots; and informs the design of future social robots that are adapted to human expectations and can effectively communicate and collaborate with people.

The current project investigates to what extent people grant robots the capacity of seeing the world from a particular point of view—of having a *visual perspective* on the world. **Do people spontaneously take a robot's visual perspective**? If so, does the same mechanism guide people's perspective taking of both robots and other humans?

We report the results of the first systematic investigation of *whether*, and *when*, people take a robot's visual perspective. To motivate our inquiry, we first visit previous research on visual perspective taking (VPT) in both human-human interaction and human-robot interaction (HRI) and develop our research questions and specific hypotheses. Then we lay out a new experimental paradigm and report three experiments on people's spontaneous (i.e., involuntary and unprompted) acts of taking a robot's visual perspective. Finding that a proportion of people indeed adopt a robot's perspective and do so under identifiable conditions, we discuss what insights these findings provide into people's perception of robots and inclination to communicate and collaborate with them.

II. BACKGROUND

A. VPT in Human-Human Interaction

Humans have remarkable abilities to simulate or infer other people's mental states, such as what they perceive, want, and feel [2, 3]. This achievement is most impressive when those mental states are very different from one's own. Being able to override an egocentric view and represent the world from another vantage point is essential to almost every aspect of human interaction, from communication, to collaboration, to empathy, and possible acts of altruism [4].

Visual perspective taking (VPT)—literally seeing the world through another's eyes—is a perceptual form of perspective taking that lays the foundation for many higher-level forms of social reasoning [5]. No two people's physical locations are alike, yet VPT helps people overcome the challenges of divergent viewpoints. By "walking in another person's shoes" and representing their unique visual experiences, people come to identify shared knowledge, reduce communication ambiguity, and achieve successful collaboration [6–8].

Although VPT plays a critical role in social cognition and interaction, people's perspective-taking abilities are far from perfect. Human infants start with a highly egocentric view of the world [9] and, step by step, start to grasp other people's divergent perspectives [10]. They first develop an appreciation that an object visible to themselves might not be visible to another observer (i.e. *what* we see is different)—an ability also known as Level-1 VPT [11]. Then, 4- to 5-year-old children start to realize that the same object(s) might appear differently to different people (i.e. *how* we see is different)—an ability also known as Level-2 VPT [12]. Even though people's perspective-taking competencies improve over the course of

development, they never completely outgrow their egocentric tendencies: even adults still make egocentric errors, and correcting those initial impulses requires time and effort [13].

Even though egocentric perspectives are natural and may often have primacy, research shows that, under certain circumstances, people reliably take others' distinct perspectives into account, such as to support dyadic interaction [14, 15]. More impressively, VPT may even occur spontaneously without any demands to communicate with the other person; it can be triggered upon merely observing another person's presence or their interaction with the environment [16–18].

B. VPT in Human-Robot Interaction

Because VPT plays such a fundamental role in social cognition and interaction, HRI researchers have shown increasing interest in enabling *robots* to take a human partner's perspective. They have explored what architecture, strategies, and designs robots should have to handle possible perspective ambiguities [20, 21], and how a perspective-taking robot might produce better interaction [22, 23]. For example, equipped with perspective-taking algorithms, robots could take into account information about a teacher's visual perspective to learn a new task [24], correctly identify the referent indicated by a human partner [25], and better identify another robot's action [26].

To the best of our knowledge, no research has examined the opposite question—whether people acknowledge a *robot's* visual perspective. The literature does show that people readily attribute other types of mental states to robots. For example, participants attributed negative emotions to a robot dinosaur when watching it being tortured [27, 28], and they attributed desires and intentions to a humanoid robot when playing a strategic game with it [29]. Moreover, in both studies, people's mentalizing-related brain regions were activated as shown by functional magnetic resonance imaging, further supporting that people attributed mental states to robots—in fact, people's anthropomorphization system is so powerful that they even ascribed minds to moving geometric shapes on a screen [30].

However, in many of the above cases, people may project their own mental states onto other agents. By contrast, in the case of perspective taking, people would need to override their own egocentric viewpoint. Considering that people often fail to appreciate other *humans'* distinct viewpoints, will people at all take a *robot's* visual perspective? There is indirect evidence that they do. Prior research showed that 18-month-old infants followed a robot's gaze to an external target [31], and adults, too, followed a robot's gaze to disambiguate its communicative references [32]. Because gaze following serves as a foundation for understanding others' perspectives [33], it stands to reason that people may spontaneously take robots' visual perspectives.

C. What Triggers the VPT Mechanism?

If people do spontaneously take a robot's visual perspective, under what conditions are they more likely to do so? Previous research has shown that instances of VPT are sensitive to certain "trigger behaviors": upon observing another person's gaze and reaching, people are more likely to take that person's perspective to report spatial relations between object or the identity of an object [16, 17]. When a cognitive response can be reliably and systematically elicited by well-defined stimuli, we have evidence for a circumscribed *mechanism*—a piece of cognitive machinery that is sensitive to certain inputs and computes predictable outputs. In the present investigation, we will not only test whether people take a robot's visual perspective at all, but also whether the same VPT mechanism will be activated when people observe both human' and robots' behaviors—that is, whether people's VPT for a robot target is sensitive to the same trigger behaviors (e.g. gaze and reaching) as their VPT for a human target. Given research showing that human observers represent robot actions in a similar manner as human actions [34], we hypothesized trigger behaviors displayed by robots may effectively activate VPT as well.

If both a robot and a human can activate the same VPT mechanism in human observers, we might still suspect that robots and humans differ in their strength of activating the VPT mechanism. Previous research has shown that people experienced greater distress over another human's plight than over a robot dinosaur's plight [28]. Likewise, people might also be somewhat less inclined to take a robot's visual perspective than to take another person's visual perspective.

Another factor that might moderate people's VPT tendency toward robots is the robot's physical appearance, especially its human-likeness. Previous research found that robots' different levels of resemblance to a human provoked a variety of effects on people's perceptions of robots [35–37]. Therefore, our project includes two robots differing in their human-likeness to test the influence of robots' appearance on VPT [35].

III. EXPERIMENTAL PARADIGM

As a first step toward identifying the trigger behaviors that would enable a robot to elicit spontaneous VPT in humans, we developed an experimental paradigm that strips away, for now, the complications of context, interaction, communication, and relationship between robots and humans. In this paradigm, we showed each participant a photo or a video depicting an agent behind a table. On the table there was a "9" from participants' own perspective, but it could also be read as a "6" from the agent's visual perspective, and participants responded to an open-ended question "*What number is on the table?*" Clearly, identifying a number "9" from another agent's perspective as "6"—without any demands to communicate with that agent—would serve as an impressive marker of spontaneous VPT. Similar verbal description tasks have been used in previous studies to investigate people's sensitivity to multiple spatial frames of reference [16–18], and the "reporting 6/9" task has been employed to capture the more complex Level-2 VPT [19].

This straightforward yet powerful paradigm allowed us to focus on tight experimental control of robots' physical appearance and behaviors that may influence the activation of spontaneous VPT. We adopted it as a between-subjects, single-trial, free-response task in all experiments in this paper, so we first introduce the information shared by all studies in this paradigm section and later describe specific variations in following experiments.

A. Participants, Procedure, and Stimuli

We recruited all our participants via Amazon Mechanical Turk (MTurk), a crowdsourcing platform commonly used as a reliable source of participants for experimental research.

Among many advantages MTurk possesses, the demographics of its workers are more representative than typical university-based research [38]. To preclude participants from taking studies in the same project twice, we also used TurkGate [39] to effectively control and verify MTurk workers' access.

In every experiment, naïve participants were randomly assigned to one of the experimental conditions. After typing in their identification codes, they opened the experiment webpage which displayed the photo/video stimulus, the "*What number is on the table?*" question, and a textbox below. The question and textbox were presented simultaneously with the photograph or immediately after the video clip, which froze on the last frame. Participants were asked to type their answers into the textbox and then click "continue" to submit them. Finally, they filled out three basic demographic questions (gender, age, and English proficiency) and received their payment codes.

B. Independent and Dependent Variables

The primary factor we manipulated in each experiment was the *trigger behaviors* the agent demonstrated. Basic conditions included: (1) the agent being merely present in the scene, looking away from the object (*presence*); (2) the agent gazing at the object (*gaze*); or (3) the agent reaching for while gazing at the object (*reaching*). (See Figure 1) In some cases we also included a control condition where nothing was behind the table (*absence*).

Another important independent variable across experiments was *agent type*: a human male, or humanoid robots such as *Aldebaran's* Nao or *Rethink Robotics's* Baxter.

To measure the degree to which people take the agent's visual perspective, we calculated the percentage of participants who answered from the agent's perspective (i.e. "6" or "six") in each condition. We refer to these percentages as "spontaneous VPT rate" throughout the paper. Because we were interested in whether people took the robot's or their own perspectives, we excluded those few participants from data analysis who provided multiple perspectives. In addition, we measured the response times (RTs) people took to generate their answers. They were measured as the duration between the moment when the webpage loaded the question and the moment when the participant submitted their answer. In the current article, we use RTs only as an exclusion criterion: Participants who took excessively long time to respond to the question (3 standard deviations beyond the average RTs in their respective conditions) were removed from data analysis—a common practice in psychological research with RTs [40].

Considering that people's VPT response was a dichotomous choice (saying "6" or "9"), we performed logit analysis in all experiments using SPSS's LOGLINEAR module. This allowed us to simultaneously estimate both main effects and interactions of the independent variables and to test specified contrasts for independent variables with more than two levels. All *p*-values reported in this paper are two-tailed. We calculated effect sizes, where possible, using Cohen's *d*.

IV. EXPERIMENT 1

Would people ever represent an object from a robot's perceptual viewpoint? If so, can robots and humans activate the same VPT mechanism in human observers to the same level? Experiment 1 took an initial step to address these questions. We used photographs as stimuli to provide parsimonious information about the agents and the behaviors.

A. Methods

Experiment 1 was a 2 (agent types) × 3 (trigger behaviors) between-subjects design with two additional control conditions. For the agent type factor, we created photographs of either a human or a humanoid robot agent sitting behind a table. The human was a white male in his 20s. The humanoid robot was a 58-cm tall red-colored Nao robot (*Aldebaran Robotics*). Importantly, because Nao does not have white sclera (which plays an essential role in gaze following and joint attention in human interaction [41]), we edited its pupils in Adobe Photoshop CS6 to make its eyes indicative of gaze direction as human eyes do (see Figure 1b).

The second manipulated factor was the trigger behaviors, which included presence, gaze, and reaching. In addition, we included two control conditions: The novelty-control condition measured people's spontaneous perspective taking when encountering a novel artifact (a colorful electric guitar) "sitting" across a table; the absence-control condition determined the baseline of spontaneous VPT with no agent or chair present in the scene.

To decide on appropriate sample sizes, we followed previous studies of similar paradigms [16] and aimed at $n = 60$ in the human condition. Unsure about the effect sizes in the robot condition, we started with a generous estimation of sufficient sample size and aimed at $n = 90$. Despite that we have preset our sample sizes, the final counts of participants in each condition usually had small variances because of 1) randomness associated with the method of list assignment on the Internet, and 2) removing participants whose RTs were 3 SDs beyond the means.

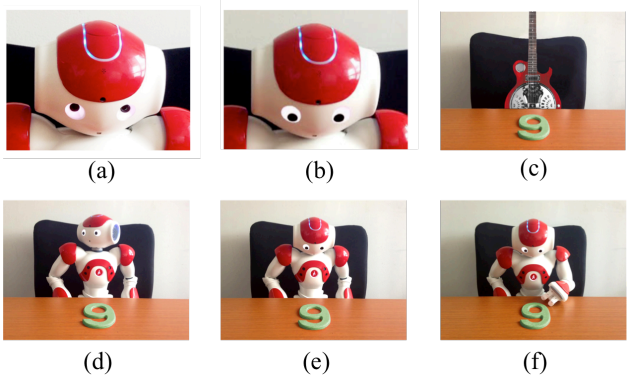

Figure 1. Nao in Experiment 1: (a)–(b) original and photo-edited eyes, (c) the novelty-control condition, (d)–(f) presence, gaze, and reaching conditions.

B. Results

In the human condition, 12 participants were removed from final analysis after RT screening; all of the remaining 236 participants (mean age = 30 years, 46% female; n = 57-63 per behavior condition) answered either "6" or "9". In the robot condition, 27 participants were removed after RT screening,

and 1 was removed due to reporting multiple perspectives, ending up with 364 participants (mean age = 33 years, 57% females, n = 86-92 per condition). Note that because we had no hypotheses related to gender or age and also found no main effects in our exploratory analyses in all three experiments, we collapsed across gender and age for final reported analyses.

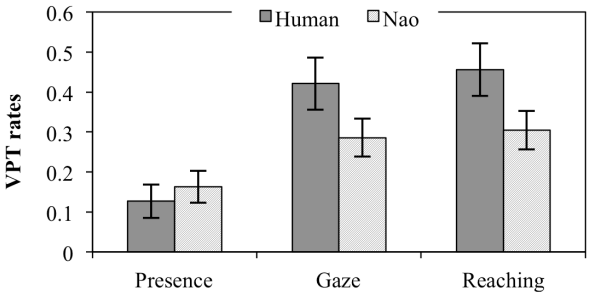

Figure 2. Spontaneous VPT rates with Nao and human agents across trigger behaviors in Experiment 1. Error bars represent +/- 1 standard error of the mean in all graphs.

The low VPT rates in control conditions confirmed that people did not have a blind tendency reporting "6" when there was either no agent (1.7%) or a novel object without perception capacity (8.3%). Logit analyses with planned contrasts showed that Nao's three trigger behaviors (presence, gaze, reaching) induced significantly higher VPT rates than the absence-control condition, z = 3.04, p = .002, d = .44, and the novelty-control condition, z = 3.20, p = .001, d = .38. Then we preformed a 2 (agent type) × 3 (trigger behaviors) logit analysis with Helmert contrasts on trigger behaviors and found a strong main effect of trigger behaviors: among three trigger behaviors, the average of gaze and reaching induced a significantly higher VPT rates than the presence, z = 4.52, p < .001, d = .71, whereas VPT rates induced by gaze and reaching did not differ from each other. We found no significant main effect of agent—Nao and the human actor triggered similar VPT rates in participants. However, the interaction between agent and the specific contrast of presence vs. gaze and reaching was marginally significant, z = 1.7, p = .09, d = .16.

C. Discussion

In Experiment 1, we found that people indeed took a humanoid robot's visual perspective, especially when it displayed object-directed gaze and goal-directed reaching, which were also the trigger behaviors that facilitated VPT in response to a human agent. There was a (marginally significant) trend that people might be more apt to take another human's visual perspective than a robot's visual perspective in response to gaze and reaching compared to mere presence.

Experiment 1 did not clarify whether the two triggering behaviors, gaze and reaching, independently activated VPT. Because agents in the reaching condition were also gazing at the object, it is possible that only gaze enhanced VPT rates (in both conditions), and reaching by itself is ineffective in creating perceptions of agency and ascriptions of point of view.

In Experiment 2 we therefore used a robot model that allowed us to manipulate reaching behavior independently from its gaze. We did so by displaying or not displaying the robot's eyes while it was reaching for the target number. Moreover, because this robot model (Baxter by *Rethink Robotics*) was less humanoid than Nao, we could test whether the patterns of VPT activation in Experiment 1 replicated with a less human-like robot.

V. EXPERIMENT 2

A. Methods

1) Stimuli

Baxter is an industrial robot built by *Rethink Robotics*. It is about 6ft tall and has a small screen toward the top that can present animated images of its eyes, which creates an impression that Baxter is gazing in a certain direction. Alternatively, the screen can be turned off, making Baxter an eyeless robot—an ideal robot agent for our purpose. In addition, Baxter features two strong arms that were designed to perform simple industrial jobs such as loading, sorting, and handling of materials on a production line, so its reaching behavior would appear unambiguous and convincing in photographs. All taken together, Baxter clearly has some resemblance to a human body, but it was rated as less human-like and more machine-like than Nao according to participants who viewed both of their photos [35].

Because the specific model available to us did not come with ready-made animated images of eyes, we used Adobe Illustrator to draw our own versions that closely resembled those designed by Rethink Robotics (see Figure 3a, 3b). To make Baxter's facial features salient in the photographs, we used Adobe Photoshop CS6 to superimpose the images of eyes on Baxter's screen region during post-production.

Each photograph was set to a 360 × 360 resolution, which was decided based on two considerations: First, Baxter's physical features should be clearly visible; second, both the visual stimuli and the question *"What number is on the table?"* could appear on the screen simultaneously without scrolling up and down the experiment webpage.

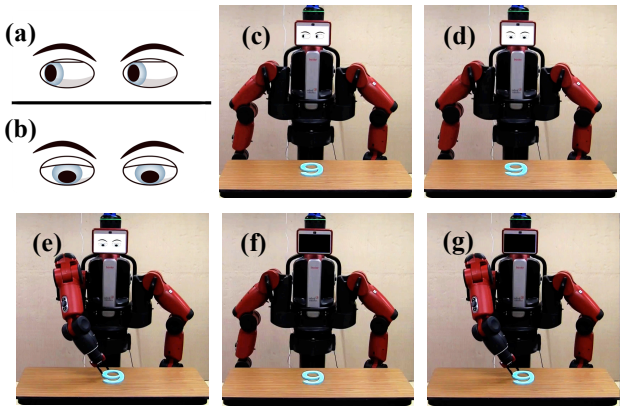

Figure 3. Baxter in Experiment 2: (a)–(b) eyes in the presence and gaze conditions, respectively; (c)–(e) presence, gaze, and reaching conditions; (f)–(g) presence and gaze without eyes.

2) Design

Experiment 2 was a between-subjects design with five experimental conditions: Presence, gaze, and reaching, all with eyes, served as conceptual replications of those in Experiment

1 (see Figure 3c-3e). In two new conditions—"presence-no eyes" and "reaching-no eyes"—Baxter's screen was turned off (see Figure 3f, 3g). Comparing VPT rates between the latter two conditions allowed us to test whether reaching by itself, without gaze, could activate VPT. In addition, we included a control condition where Baxter was entirely absent, replicating the absence-control condition in Experiment 1.

3) Participants

After getting an initial sense of people's VPT responses to a robot, we were confident that our robot study could use a typical sample size of 60. Therefore, after taking randomness and unqualified data into account, we set $n = 70$ in the data-collection system. Experiment 3 followed the same rule.

B. Results

Twenty-six participants were excluded after RT screening, and 6 more were excluded because of reporting multiple perspectives, yielding 388 participants for data analysis (mean age = 32 years, 61.0% female, n = 60-68 per condition).

The VPT rate of the absence-control condition (1.7%) replicated that of the same condition in Experiment 1 (also 1.7%). The left panel of Figure 4 depicts VPT rates across the three trigger behaviors shared with Experiment 1, while the right panel of Figure 4 depicts VPT rates of the two newly added no-eyes conditions. We first analyzed the conditions in the left panel (along with the control condition) using logit analysis with Helmert contrasts. Once again, the three trigger behaviors (presence, gaze, reaching) induced significantly higher VPT rates than the absence-control condition, $z = 2.73$, $p = .006$, $d = .40$. Among those behaviors, the average of gaze (22.4%) and reaching (24.3%) somewhat increased VPT rates from presence (15.9%), but this difference did not reach statistical significance, $z = 1.11$, $p = .27$, $d = .17$. VPT rates in the gaze condition and the reaching condition also did not significantly differ from each other.

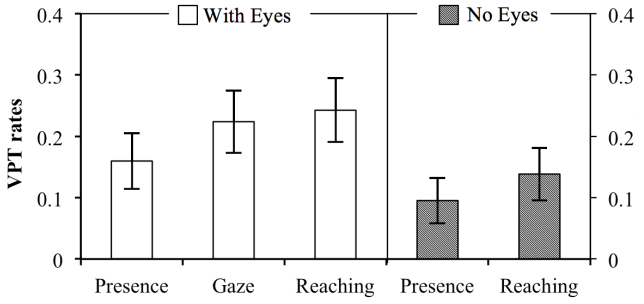

Figure 4. Spontaneous VPT rates with Baxter across trigger behaviors in Experiment 2. Left panel: all three behaviors with eyes; right panel: presence & reaching without eyes.

The right panel where eyes were removed illustrated that most people did not take Baxter's visual perspective when it was merely present—the VPT rate of 9.5% was almost as low as that in the novelty-control condition in Experiment 1 (8.3%). Then we examined whether reaching by itself, without gaze, induced stronger VPT tendency, and we found that it barely increased VPT rates relative to mere presence, $z = .73$, $p = .46$. Meanwhile, when gaze was added back to reaching (i.e. comparing reaching across panels), we saw a suggestive, yet insignificant increase in VPT rates, $z = 1.20$, $p = .23$, $d = .21$.

C. Discussion

Using a less human-like robot, we largely replicated the finding in Experiment 1 that human observers took a robot's visual perspective to a moderate level. However, different from results in Experiment 1, in the current study, gaze and reaching were only marginally more effective in triggering VPT than presence. This might be partially explained by the decreased statistical power in Experiment 2 as a result of lower VPT rates in the gaze and reaching conditions and a reduced sample size.

Despite suggestive statistics, we did not find evidence that a robot's reaching, without the company of gaze, activated more VPT than a robot's mere presence. This result seems surprising in light of previous research where reaching, when performed by a human hand [42, 43]—or occasionally a robotic hand [34]—triggered goal representations in observers even without the presentation of eyes. Assuming that a robot's reaching could elicit a reasonably strong perception of goal and agency, why weren't people more willing to adopt the robot's visual perspective in the reaching condition?

One possible explanation is that, in Experiments 1 and 2, people observed only still images of reaching behavior, which could not sufficiently trigger underlying goal representations in human observers in the first place. In the majority of previous research investigating such effects, the visual stimuli were either videos or live demonstrations instead of still images. This may suggest that observing behaviors in motion provides reaching with greater validity. Moreover, in their daily lives people rarely see robots interacting with objects, so they might fail to interpret the photos of a robot's unfamiliar hand gestures as reaching at all, therefore not attributing goals to this specific hand-object configuration. For example, participants inspecting the photos might have believed that the agent positioned its gripper for other reasons, or the agent was simply moving its gripper without purpose and accidently placed one close to the number. The fact that our robot's reaching behavior did not include any finger/gripper-grasping movement might have further contributed to the ambiguity of reaching poses.

For Experiment 3, we decided to use videos instead of still images to display a robot's behaviors. We expected that this more dynamic format might disproportionately benefit the reaching condition's ability to trigger spontaneous VPT.

VI. EXPERIMENT 3

A. Method

We examined a 3 (agent types: human, humanoid robot, mechanical robot) × 3 (trigger behaviors: presence, reaching, gaze) between-subjects factorial design. In addition, we also included the "reaching-no eyes" condition with Baxter; this condition allowed us to investigate whether goal-directed reaching without eyes, when presented in motion, could effectively trigger spontaneous VPT—an unresolved question from Experiment 2. (No new "presence-no eyes" condition was included because, without motion, it would be identical to the already existing photo condition in Experiment 2. We refer to the data collected in Experiment 2 where appropriate.)

We designed five-second video recordings as the visual representations of all three agents. We recorded these videos using a Canon HD VIXIA HFS100 video camera. We employed Nao's and Baxter's movement-recording interface to make their head-turning and hand-reaching movements smooth and natural. To make sure that videos of the same condition were maximally similar across agents, all videos were produced under similar protocols of movement sequences.

Specifically, for the presence condition, both human actor and Nao started by looking into the video camera for approximately 2-2.25 seconds, then turned their heads to look to the right side; the head turn took approximately 1 second. For Baxter, because we could not turn its screen to the side, we created a "stop motion" video with four animated pictures of its eyes: eyes closed (0.5s); looking to the front (1.75s); intermediately to the right (0.25s), and completely to the right (2.5s) (see Figure 3a).

For the gaze condition, the beginning of the videos was identical to the presence condition. Then, both the human actor and Nao turned their heads down to look at the number. The head turn again took approximately 1 second to finish. Because we could not lower Baxter's screen, we presented three pictures of its eyes on the screen: eyes closed (0.5s), looking to the front (2.5s), and looking down (2s) (see Figure 3b).

For the reaching condition, both the human actor and Nao started by looking into the video camera for approximately 1.5 seconds, then they started to lower the heads and reached their hands from below the table when their heads were half down. The entire gaze-reaching sequence took approximately 3 seconds. For Baxter, its hand started to move approximately 0.25 second after the start of the video. The entire reaching took about 4 seconds to finish—longer than Nao's and the human's reaching movements because Baxter's arms were longer and travelled more slowly. In the case where the screen was turned on and Baxter's eyes were displayed, it maintained "eye contact" for 0.75 seconds, then started the gaze sequence.

We used Apple Final Cut Pro X to adjust video speed, add color contrast, and reduce the resolution to 360 × 360—the video size presented to participants on the experiment webpage. Similar to Experiment 2, we also superimposed images of eyes onto Baxter's screen region for the eyes-present videos (using Final Cut Pro X), because the contents of Baxter's small screen appeared blurry in the original videos.

In every experimental condition, the video clip "froze" on it last frame on the webpage. Then, participants received the critical response prompt below the frame on the same page.

B. Results

Twenty-eight participants were excluded after RT screening, and 3 more were excluded because their responses contained multiple perspectives, yielding 660 participants for data analysis (mean age = 33 years, 57.6% females, n = 63-70 per condition). Among them, 594 participants were randomly assigned to the 3 (agent types) × 3 (trigger behaviors) design. A 3 × 3 logit analysis with Helmert contrasts on both factors revealed that people were significantly more inclined to take the human actor's perspective than to take the two robots' perspectives, z = 3.15, p = .002, d = .27, and there was no difference between the robots, z = .131. We also found that, compared with VPT rates in the presence condition, gaze and reaching elicited significantly higher VPT rates, z = 4.41, p < .001, d = .38; moreover, reaching evoked significantly stronger VPT than gaze, z = 2.53, p = .01, d = .25 (see Fig. 5). No significant interaction effect was found.

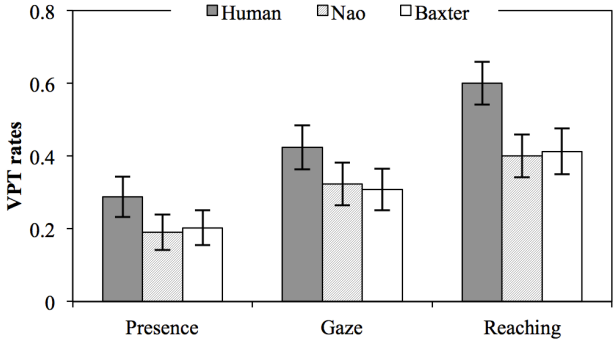

Figure 5. Spontaneous VPT rates with human, Nao, and Baxter across three trigger behaviors in Experiment 3.

In addition, 66 participants responded to the "reaching-no eyes" condition, yielding a VPT rate of 30.3%—significantly higher from that in the "presence-no eyes" condition in Experiment 2 (9.5%), z = 2.77, p = .006, d = .49, yet marginally lower than the reaching condition with gaze (41.3%), z = 1.64, p = .101, d = .29.

Finally, we conducted cross-sample analyses of Experiments 1, 2 and 3 to answer the following two questions: Did VPT rates differ in response to human and robot agents? Were people more likely to take another agent's perspective when they saw that agent in motion rather than in a still image?

C. Cross-Experiment Analyses

We selected three conditions commonly shared by all three experiments, namely presence, gaze, and reaching (with eyes) and conducted a 3 (agent: human, Nao, Baxter) × 2 (medium: picture, video) × 3 (trigger behaviors: presence, gaze, reaching) logit analysis, with Helmert contrasts on both agent type and trigger behaviors. We found that, overall, videos indeed induced significantly higher levels of spontaneous VPT than photographs, z = 3.52, p < .001, d = .20. We also found that the human agent invited significantly higher VPT rates than the robot agents, z = 3.61, p < .001, d =.22. By contrast, VPT rates elicited by the two robots did not differ from each other, z = .66, p = .51. Across agents, we also found that gaze and reaching were more effective in triggering VPT than mere presence, z = 6.14, p < .001, d = .37, and reaching was more effective than gaze, z = 2.01, p = .022, d = .14. Of four interaction contrasts, one reached statistical significance, suggesting that increases of VPT rates from presence to gaze or reaching were stronger when elicited by human agents than by robot agents, z = 1.786, p = .037, d = .11. This pattern mirrored the interaction effect reported in Experiment 1.

VII. GENERAL DISCUSSION

This project is a first attempt to examine whether, and when, human observers spontaneously take a robot's visual perspective and identify an object from its vantage point. In

three studies, we compared three potential triggering behaviors (i.e., *presence* without engagement with the object, object-directed *gaze*, and goal-directed *reaching*) performed by two robot models that differed in their human-likeness. We found that a notable proportion of people took the robots' perspectives upon observing their gaze and reaching behaviors. Furthermore, reaching was more effective in activating VPT than gaze when these triggering behaviors were presented on videos instead of on photos. Even though VPT rates in response to robot agents were lower overall than those in response to human agents, their *patterns* were highly similar. This suggests that robots elicit the same cognitive mechanism of VPT that we know is elicited by a human agent.

The current research sheds light on how people perceive robots and on some of the downstream consequences of such perception. Recognizing that a seemingly obvious "9" can be a "6" from a robot's viewpoint requires some implicit appreciation that a robot may hold distinct representations. Our task therefore indirectly measures people's acceptance of robots as social agents with subjective internal states rather than mere automata. Our results also extend previous research on the cognitive effects of observing robot actions. For instance, while [34] showed that robots' reaching or grasping gestures subtly (by a few milliseconds) influenced people's reaction times in an irrelevant subsequent visual search task, our task demonstrated that, upon brief exposure to robots' gaze and reaching behaviors, people overrode the primacy of their egocentric viewpoints and explicitly adopted the robot's conflicting viewpoint. Because the current research uniquely incorporates examination of language use (selecting a viewpoint in object descriptions) with research on action representation and mind perception, we believe it could have further relevance to research on human-robot communication.

In addition to revealing the role a robot's nonverbal behaviors plays in eliciting spontaneous VPT, our results also inform how robots' overall physical features, such as human-likeness, may influence people's spontaneous VPT tendency. Although we have shown that people were less inclined to take robots' perspectives than taking a person's perspective, we have not found differences between two robots that differ moderately in human-likeness. We are currently investigating whether such similarity in VPT responses generalizes to robots that are even more human-like, especially those deep in the "uncanny valley" [35], where substantial yet imperfect human-likeness often elicits discomfort in human observers. To our knowledge, only one study has investigated whether computer-animated characters' physical appearance moderates spontaneous Level-1 VPT [45], and researchers found no consistent relationship between the human-likeness or eeriness of the characters and the activation of Level-1 VPT. However, because Level-2 VPT requires more mentalistic attribution than Level-1 VPT, it remains an open question whether an android (nearly indistinguishable from human appearance) may elicit stronger spontaneous VPT than Nao and Baxter or whether its eeriness—especially when in motion—will backfire.

A clear limitation of the present studies is that our paradigm is devoid of actual interaction between participants and the agent whose perspective they may take. Although we intentionally chose to use photographs and video recordings because of the tighter experimental control they afford, we cannot predict, without further empirical evidence, to what extent the same effects would also apply to real-time interaction. Nevertheless, based on the increased VPT rates from the photo to the video condition, we speculate that watching a robot perform trigger behaviors "in person" is likely to further increase VPT. As a next step, it would therefore be fruitful to explore spontaneous VPT in more interactive paradigms. For example, one might employ joint tasks adapted from [13] in which human speakers provide verbal instructions to another human or robot partner and thereby reveal the perspectives they take. In particular, such tasks could reveal to what extent speakers adjust away from their own viewpoint and recognize their partner's unique view.

This limitation notwithstanding, our research has important implications for the design of social robots. Humans and robots will, in the near future, collaborate on joint tasks that often demand visual and spatial alignment, so resolving the issue of divergent perspectives is critical to achieving effective HRI. Our research suggests that humans are indeed inclined, under certain circumstances, to engage with a robot as a collaborator with a "viewpoint." Based on our findings, if designers want to elicit such spontaneous engagement, they can equip their robots with the right trigger behaviors; if designers want to circumvent such spontaneous engagement, however, our project suggests which trigger stimuli to avoid.

Furthermore, our study indicates new directions in perspective taking research in HRI. Current researchers and designers focus primarily on *how* to equip robots with perspective-taking capacity; largely missing has been research on *when* robots should perform such capacities to achieve most intuitive and effective interaction. Although having some perspective-taking capacity is generally desirable for a robot (and also for a human), we don't know yet whether a robot's persistent perspective-taking tendencies might backfire in certain situations and undermine joint performance. Perspective taking is, after all, a constant coordination among multiple parties in response to dynamically changing situations. Therefore, future investigation on when people find it natural to take a robot's perspective, and when robots should take the lead in perspective taking, may help us achieve more natural and efficient human-robot interaction.

ACKNOWLEDGEMENT

This project was supported in part by a grant from the Office of Naval Research, No. N00014-14-l-0144. The opinions expressed here are our own and do not necessarily reflect the views of ONR.

We thank Matthias Scheutz's Human-Robot Interaction Laboratory at Tufts University and Stefanie Tellex's Humans to Robots Laboratory at Brown University for contributing Nao and Baxter, respectively, for stimulus production.

REFERENCES

[1] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, pp. 433-460, Oct. 1950.

[2] B. F. Malle, "Folk theory of mind: Conceptual foundations of human social cognition," in *The New Unconscious*, R. Hassin, J. S. Uleman,

and J. A. Barge, Eds. New York: Oxford University Press, 2005, pp. 225-255.

[3] H. Wellman, *The child's theory of mind*. Cambridge, MA: MIT Press, 1990.

[4] S. D. Hodges, B. Clark, and M. W. Myers, "Better living through perspective taking," in *Positive Psychology as a Mechanism for Social Change,* R. Biswas-Diener, Ed. Dordrecht, The Netherlands: Springer Press, 2011, pp. 193-218.

[5] J. H. Flavell, "Development of knowledge about vision," in *Thinking and Seeing: Visual Metacognition in Adults and Children,* D. T. Levin, Ed. Cambridge, MA: MIT Press, 2004, pp. 13-36.

[6] H. H. Clark, *Using language*. Cambridge, England: Cambridge University Press, 1996.

[7] D. Gergle, R. E. Kraut, and S. R. Fussell, "Using visual information for grounding and awareness in collaborative tasks," *Hum.–Comput. Interact.*, vol. 28, no. 1, pp. 1-39, 2013.

[8] S. E. Brennan, A. Galati, and A. K. Kuhlen, "Two minds, one dialog: coordinating speaking and understanding," in *The Psychology of Learning and Motivation*, vol. 53, B. H. Ross, Ed. Burlington: Academic Press, 2010, pp. 301-344.

[9] J. Piaget, and B. Inhelder, B, *The child's conception of space*. London: Routledge & Kegan Paul, 1956.

[10] Z. S. Masangkay, K. A. McCluskey, C. W. McIntyre, J. Sims-Knight, B.E. Vaughn, and J. H. Flavell, "*The early development of inferences about the visual percepts of others*," *Child. Dev.*, vol. 45, no. 2, pp. 357-366, Jun. 1974.

[11] H. Moll and M. Tomasello, "Level 1 perspective-taking at 24 months of age," *Brit. J. of Dev. Psychol.*, vol. 24, no. 3, pp. 603-613, Sep. 2006.

[12] H. Moll, A. Meltzoff, K. Merzsch and M. Tomasello, "Taking versus confronting visual perspectives in preschool children," *Dev. Psychol.*, vol. 49, no. 4, pp. 646-654, Apr. 2013.

[13] N. Epley, C. Morewedge and B. Keysar, "Perspective taking in children and adults: Equivalent egocentrism but differential correction," *J. Exp. Soc. Psychol.*, vol. 40, no. 6, pp. 760-768, Sep. 2004.

[14] M. Schober, "Spatial perspective-taking in conversation," *Cognition*, vol. 47, no. 1, pp. 1-24, Apr. 1993.

[15] A. Nadig and J. Sedivy, "Evidence of Perspective-Taking Constraints in Children's On-Line Reference Resolution," *Psychol. Sci.*, vol. 13, no. 4, pp. 329-336, Jul. 2002.

[16] B. Tversky, and B. M. Hard, Embodied and disembodied cognition: Spatial perspective-taking, *Cognition*, vol. 110, no. 1, pp. 124–129, Jan 2009.

[17] X. Zhao, C. Cusimano, and B. F. Malle, In Search of Triggering Conditions for Spontaneous Visual Perspective Taking, in *Proc. 37th Annual Conf. of Cognitive Science Society*, Los Angeles, CA, 2015, pp. 2811-2816.

[18] A. D. Surtees, M. L. Noordzij and I. A. Apperly, "Sometimes losing your self in space: Children's and adults' spontaneous use of multiple spatial reference frames," *Dev. Psychol.*, vol. 48, no. 1, pp. 185-191, Jan. 2012.

[19] A. D. Surtees, S. A. Butterfill and I. A. Apperly, "Direct and indirect measures of Level-2 perspective-taking in children and adults," *Brit. J. of Dev. Psychol.*, vol. 30, no. 1, pp. 75-86, Mar. 2012.

[20] M. Berlin, J. Gray, A. L. Thomaz, and C. Breazeal, "Perspective taking: An organizing principle for learning in human-robot interaction," in *Proc. AAAI*, 2006, pp. 1444-1450.

[21] E. Sisbot, R. Ros, and R. Alami. "Situation assessment for human-robot interaction," In *RO-MAN, 2011 IEEE*, Atlanta, GA, 2011, pp. 15-20.

[22] C. Torrey, S. R. Fussell, and S. Kiesler, "What robots could teach us about perspective taking," in *Expressing Oneself/Expressing One's Self: A Festschrift in Honor of Robert M. Krauss,* E. Morsella, Ed. New York: Taylor and Francis, 2010, pp. 93-106.

[23] J. G. Trafton, A. C. Schultz, M. Bugajska, F. Mintz, "Perspective-taking with robots: experiments and models," in *RO-MAN,* 2005, pp. 580-584.

[24] C. Breazeal, M. Berlin, A. Brooks, J. Gray and A. Thomaz, "Using perspective taking to learn from ambiguous demonstrations," *Robot. Auton. Syst.*, vol. 54, no. 5, pp. 385-393, May 2006.

[25] J. G. Trafton, N. L. Cassimatis, M. D. Bugajska, D. P. Brock, F. E. Mintz, and A. C. Schultz, "Enabling effective human–robot interaction using perspective-taking in robots," *IEEE T. Syst. Man Cy. A*, vol. 35, no. 4, pp. 460–470, 2005.

[26] M. Johnson and Y. Demiris, "Perceptual Perspective Taking and Action Recognition," *Int. J. Adv. Robot. Syst.*, vol. 2, no. 4, 301-308, Apr. 2005.

[27] A. M. Rosenthal-von der Pütten, N. C. Krämer, L. Hoffmann, S. Sobieraj, and S. C. Eimler, "An experimental study on emotional reactions towards a robot," *Int. J. Soc. Robot.*, vol 5, no. 1, pp. 17-34, Jan. 2013.

[28] A. M. Rosenthal-von der Pütten, F. P. Schulte, S. C. Eimler, S. Sobieraj, L. Hoffmann, S. Maderwald, M. Brand, and N. C. Krämer, "Investigations on empathy towards humans and robots using fMRI," *Comput. Hum. Behav.,* vol. 33, pp. 201-212, Apr. 2014.

[29] S. Krach, F. Hegel, B. Wrede, G. Sagerer, F. Binkofski and T. Kircher, "Can Machines Think? Interaction and Perspective Taking with Robots Investigated via fMRI," *PLoS ONE*, vol. 3, no. 7, p. e2597, Jul. 2008.

[30] F. Heider and M. Simmel, "An experimental study of apparent behavior," *Am. J. Psychol.*, vol. 57, no. 2, pp. 243-259, Apr. 1944.

[31] A. Meltzoff, R. Brooks, A. Shon and R. Rao, " "Social" robots are psychological agents for infants: A test of gaze following," *Neural Networks*, vol. 23, no. 8-9, pp. 966-972, Oct.-Nov. 2010.

[32] M. Staudte and M. Crocker, "Investigating joint attention mechanisms through spoken human–robot interaction," *Cognition*, vol. 120, no. 2, pp. 268-291, Aug. 2011.

[33] H. Moll, and A. N. Meltzoff, "Joint attention as the fundamental basis of perspectives," in *Joint Attention: New Developments in Psychology, Philosophy of Mind, and Social Neuroscience*, A. Seemann, Ed. Cambridge, MA: MIT Press, 2011, pp. 393-413.

[34] A. Wykowska, R. Chellali, M. M. Al-Amin and H. J. Müller, "Implications of Robot Actions for Human Perception. How Do We Represent Actions of the Observed Robots?", *Int. J. Soc. Robot*., vol. 6, no. 3, pp. 357-366, Jul. 2014.

[35] M. Mathur and D. Reichling, "Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley," *Cognition*, vol. 146, pp. 22-32, Jan. 2016.

[36] F. Eyssel, D. Kuchenbrandt, S. Bobinger, L. de Ruiter, and F. Hegel, "'If you sound like me, you must be more human': On the interplay of robot and user features on human-robot acceptance and anthropomorphism," *In Proc. Human-Robot Interaction*, Portlan, OR, 2012, pp. 125–126.

[37] P. J. Hinds, T. L. Roberts, and H. Jones, "Whose job is it anyway? A study of human-robot interaction in a collaborative task," *Hum.-Comput. Interact.*, vol. 19, no. 1–2, pp. 151–181, Mar. 2004.

[38] G. Paolacci, J. Chandler, and P. G. Ipeirotis, "Running experiments on amazon mechanical turk," *Judgm. Decis. Mak.*, vol. 5, no. 5, pp. 411-419, Aug. 2010.

[39] GitHub, "gideongoldin/TurkGate", 2015. [Online]. Available: https://github.com/gideongoldin/TurkGate. [Accessed: 14- Dec- 2015].

[40] R. Zwaan, and D. Pecher, "Revisiting Mental Simulation in Language Comprehension: Six Replication Attempts," *PLoS ONE,* vol. 7, no. 12, p. e51382, Dec. 2012.

[41] M. Tomasello, B. Hare, H. Lehmann, and J. Call, "Reliance on head versus eyes in the gaze following of great apes and human infants: The cooperative eye hypothesis," *J. Hum. Evol*., vol. 52, no. 3, pp. 314–320, Mar. 2007.

[42] A. L. Woodward, "Infants selectively encode the goal object of an actor's reach," *Cognition*, vol. 69, no. 1, pp. 1-34, Nov. 1998.

[43] G. Rizzolatti, L. Fogassi, and V. Gallese, "Neurophysiological mechanisms underlying the understanding and imitation of action," *Nat. Rev. Neurosci.*, vol. 2, no. 9, pp. 661–670, Sep. 2001.

[44] L. J. Byom, and B. Mutlu, "Theory of mind: mechanisms, methods, and new directions," *Front. Hum. Neurosci.*, vol. 7, article 413, Aug. 2013.

[45] K. MacDorman, P. Srinivas and H. Patel, "The uncanny valley does not interfere with level 1 visual perspective taking," *Comput. Hum. Behav*, vol. 29, no. 4, pp. 1671-1685, Jul. 2013.