Eye Movements as a Means To Evaluate and Improve Robots

Jan Zwickel
Department Psychology
Ludwig-Maximilians-University
Leopoldstraße 13
D-80802 Munich
Germany
zwickel@psy.uni-muenchen.de

Hermann J. Müller
Department Psychology
Ludwig-Maximilians-University
Leopoldstraße 13
D-80802 Munich
Germany
hmueller@psy.lmu.de

International Journal of Social Robotics manuscript No.
(will be inserted by the editor)

Eye Movements as a Means To Evaluate and Improve Robots

Jan Zwickel · Hermann J. Müller

Received: date / Accepted: date

Abstract With an increase in their capabilities, robots start to play a role in everyday settings. This necessitates a step from a robot-centered (i.e., teaching humans to adapt to robots) to a more human-centered approach (where robots integrate naturally into human activities). Achieving this will increase the effectiveness of robot usage (e.g., shortening the time required for learning), reduce errors, and increase user acceptance. Robotic camera control will play an important role for a more natural and easier-to-interpret behavior, owing to the central importance of gaze in human communication. This study is intended to provide a first step towards improving camera control by a better understanding of human gaze behavior in social situations. To this end, we registered the eye movements of humans watching different types of movies. In all movies, the same two triangles moved around in a self-propelled fashion. However, crucially, some of the movies elicited the attribution of mental states to the triangles, while others did not. This permitted us to directly distinguish eye movement patterns relating to the attribution of mental states in (perceived) social situations, from the patterns in non-social situations. We argue that a better understanding of what characterizes human gaze patterns in social situations will help shape robotic behavior, make it more natural for humans to communicate with robots, and establish joint attention (to certain objects) between humans and robots. In addition, a better understanding of human gaze in social situations will provide a measure for evaluating whether robots are perceived as social agents rather than non-intentional machines. This could help decide which behaviors a robot should display in order to be perceived as a social interaction partner.

Keywords social robots · eye movements · theory of mind

Jan Zwickel
Department Psychology
Ludwig-Maximilians-University
Leopoldstraße 13
D-80802 Munich E-mail: zwickel@psy.uni-muenchen.de

Hermann J. Müller
Department Psychology
Ludwig-Maximilians-University
Leopoldstraße 13
D-80802 Munich E-mail: hmueller@psy.lmu.de

1 Introduction

1.1 Roles of Gaze

Gaze plays a central role in human everyday life. On the one hand, it provides feedback about the success and failure of particular (motor) actions, that is, it contributes to ensuring goal achievement in online action control. On the other, gaze plays also a role in human-human interactions. In such situations, gaze is closely coupled to attention [1,2], conveys intent [3], regulates social interactions [4], and is used as a pointer to objects of shared interest [5]. Even though it is less immediately apparent that gaze is functionally necessary in this context, it is clearly socially mandatory. For example, a robot bar tender could serve a drink with camera eyes averted from the customer, but this would be inconsistent with what the customer expects in such a situation. On the other hand, fixating ("staring at") the customer all the time would be experienced as discomfort. Up to now, research has focused mostly on the first type of situation. However, arguably, the social factors in robotic camera control will come to be seen as increasingly important for more complex human-robot interaction scenarios.

The importance of gaze can be observed quite early in human development. Gaze starts to play a fundamental role in everyday life during infancy and preserves its central role during adulthood. Even 12-month old babies follow the gaze of "entities" that posses attributes typically associated with

"agency", that is, as possessing a face or showing contingent interactive behavior [6]. Infants use this gaze-following behavior to learn (to predict) where interesting events are likely to occur and computational models are available of the acquisition of gaze-following behavior [7,8]. Later in life, eye gaze cues nearly automatically attract attention to looked-at locations or objects, and lack of this behavior has been linked to disorders such as autism [9].

The social function of gaze is not restricted to increasing comfort. Rather, social gaze is also relevant when communication is needed for task achievement. Thus, for example, robot cameras mimicking human eye movements can improve communication about common goals in joint work situations, as for instance in an assembly task in which humans and robots have to communicate about the next (task) step to perform (e.g., the next object to approach etc.). Human-gaze-like behavior would be a natural way to achieve this communicative goal. However, to implement such behavior in robots, many open questions remain to be answered concerning the issue of social-gaze control in humans.

In situations such as those outlined above, human gaze control can be used as a model of robotic camera control to increase interaction efficiency. An additional advantage of a better understanding human social-gaze control is related to robot evaluation. There is as yet no agreed standard for assessing how human-like a robot is perceived [10]. Arguably, however, based on knowing how human gaze patterns change in social, as compared to non-social, situations, indices can be developed that reflect whether a robot is perceived as a social agent rather than a pre-fixed machine. Thus, our general aim to achieve an improved understanding of how gaze patterns change in social compared to non-social situations translates in two closely related goals. One is to implement the knowledge gained about social-gaze patterns in robotic camera control. The other goal is to derive indices from this knowledge that permits the human-likeness of robots to be evaluated.

Despite the relevance of gaze in dynamic situations and the many studies that examined gaze patterns in different task contexts [11–13], most studies looked at gaze pattern changes in *static* social situations. For static social situations, faces and, in particular, eyes have been shown to be strong attractors for the onlooker's gaze [11]; that is, stimulus aspects associated with social agents (such as faces) play a prominent role in determining gaze behavior. In more dynamic situations, however, certain patterns of stimulus movement unfolding in time can give rise to neutral objects being perceived as intentional agents [14]. As robots are typically employed in dynamic contexts, these are the kind of situations that need further exploration - because in such situations robots, too, can be attributed mental states, intent, and goals. This will change the expectations that humans will have of robots and how tasks are represented. For example, would humans also describe object locations relative to "intentional" robots, in the same way as they take other humans as reference points for localizing objects [15]?

In fact, humans have a tendency to attribute goals even to artifacts [16], and understanding the behavior of others as being goal-driven can be seen as a first step to developing socially meaningful interactions. Thus, the fact that humans have a preference for "teleological explanations" also with regard to the behaviors of artifacts, suggests that human roles can be taken by robots. Classical examples for attribution of social or human roles to non-human entities include the computer "therapist" ELIZA of Joseph Weizenbaum [17], which - despite its simple conversation rules - convinced some people to be a human; or the Clever Hans, a horse that was attributed arithmetic capabilities, even though it only reacted to subtle cues of its coach. A more recent demonstration is the robot Kismet [3], which - despite not being very close to human morphology - provides some of the facial features necessary for conveying emotions [18]. Moreover, even artifacts that displayed even less similarity with human morphology have been shown to evoke social interpretations [14]. For example, Heider and Simmel [14] could engender social interpretations by presenting simple geometric shapes, two moving triangles and a disc that followed different trajectories. Despite their simplicity, by virtue of their movements, these simple geometric forms gave rise to an interpretation, on the part of the onlookers, in human terms. This effect of particular movements unfolding in time is remarkable in view of the fact that simply increasing the visual similarity of artificial agents to humans does not necessarily achieve the same end. Evidence for this comes from neuro-cognitive studies.

1.2 Differential Processing of Biological and Non-Biological Information

Humans process visual information differently depending on whether it is seen as biolocial or non-biological in origin. At first, it appeared that the difference in processing was caused by differences in the visual information provided. A key finding in experimental psychology is that observing the action of others has a direct influence on motor processing, rather than just on visual perception (see, e.g., [19]). Only recently, some studies have also looked at the influence of observed artifacts' actions on human action processing. For instance, Kilner and colleagues [20] had participants perform repeated horizontal (left-right) or, respectively, vertical (up-down) movements while they watched either another human or a robot performing movements that were either congruent (the same as) or incongruent (different to) with their own movement direction. It turned out that incongruent movements of observed humans, but not of observed robots, increased the variance of the performed movements. This

was interpreted within the framework of the mirror neuron theory [21].

Mirror neurons are neurons that fire during action observation *and* execution. Some theories also relate them to social understanding [22, 23]. In the study of Kilner and colleagues, presumably only human actors, but not robots, activated the mirror neurons and therefore, in case of action incongruity, interfered with the movement actually to be performed. In the robot movements, by contrast, some biological features critical for activating the mirror neuron system seemed to have been missing. Similarly, Pozzo and colleagues [24] showed that humans could better predict the final position of a moving dot that disappeared behind an occluding surface if the dot followed a biological (rather than a non-biological) movement profile, which presumably permitted them to rely on a forward model for prediction (see also [25]).

However, in more recent studies [26, 27], even a moving dot with a non-biological movement profile has been found to interfere with movement execution. One interpretation offered by Kilner and colleagues was that the mirror system only responds if the observed stimuli combined with the observed movement patterns are sufficiently familiar - which was assumed to be the case for moving dots, but not for robots. The relevance of physical attributes of visual information is further questioned by a number of findings that point to the importance of abstract "features", in particular, the goals or interpretations, of actions for the activation of the mirror neuron system [28]. In the study of Gazzola and colleagues, a robot action was found to activate the mirror neuron system only if the observed action had a familiar goal. The predominant influence of interpretation, versus that of sheer morphological similarity, was underscored by Stanley and colleagues [29], who reported that the same moving dot stimuli did or did not lead to interference depending on whether participants were made to believe that the dot motion was biological or non-biological in origin. This also fits well with the proposal of Biocca and colleagues [30] that mental models are activated when intelligent behavior is detected. Interestingly, they linked the activation of the mental model to the feeling of social presence.

Taken together, these findings suggest that whether or not an "intentional stance" [31] is taken depends not so much on similarity to humans, but rather on other factors. Arguably, therefore, it would be advantageous to investigate human information processing in the context of dynamic social situations. An improved understanding of how human information processing differs between social and non-social situations would allow us to derive behavioral indicators for conditions in which a robot is perceived as a social agent and, based on these, tailor the robots to meet the needs of their human interaction partners.

1.3 Non-Eye Movement Measures of Intentionality Ascription

[32] had participants segment animations involving two moving stimuli into parts. Different groups of participants were given different types of information as to how the animations were created. The interpretation either suggested some intentions behind the stimuli's movements, or that the movements were randomly generated. One intentional interpretation was, for example, that the two stimuli represented the movements of two players of a chasing game. It was found that participants' segmentation decisions were more driven by bottom-up visual features in the non-intentional than in the intentional conditions. In the latter conditions, the way in which participant segmented the animations was presumably influenced by the (top-down) story information provided to them.

While these findings illustrate the influence of top-down information on (the explicit measure of) participants' segmentation decisions, these were explicitly required by the task. However, explicit measures such as this would make it hard to assess attributions of intentionality in real-world scenarios without impacting on the nature of the task. Arguably, examining observers' gaze behaviour would provide for a less intrusive way to make this assessment, which is why the present study focused on implicit eye movement measures. Further, while participants in [32] used different segmentation strategies in the different conditions, it is not really clear whether they attributed "minds" to the *stimuli*, or to the *persons* who created them. Of course, this is not a shortcoming of [32] which was designed to address a different question; however, it would be a problem with regard to the current question. Therefore, in the present study, we used rather artificial animations which permitted us to control whether or not mind attribution to the stimuli occurred.

2 Examining Eye Movements in the Context of Social Agents

That eye movements, too, are influenced by the goals of an observer has been very well documented in natural situations [33]. Furthermore, the value of rewards to be found at new fixation location has been shown to be an important factor for saccadic target selection [33, 34]. Given the dependency of eye movements on task context and reward, one would also expect to find different gaze patterns in social compared to non-social contexts.

2.1 Earlier Studies of Eye Movement Control in the Context of Social Situations

A first step in examining eye movement differences during interactions with interactive robots and non-interactive toys was taken by Dautenhahn and colleagues [35]. They measured the gaze durations of children with autism who were playing with either a robot or a (toy) truck. Unfortunately, no clear differences were found. Furthermore, eye movements are known to differ in several respects between individuals with and without autism [36]; therefore, eye movements of individuals with autism would not provide a good model for eye movement control in humans in general. Also, it is unclear what role the children attributed to the toy/robot.

[11] investigated the influence of the social content of pictures on observers' gaze behavior, where the social content was manipulated by varying the number of people depicted in different scenes. Birmingham found that, increasing the number of humans in a scene led to an increase in gazes directed to the eyes of the humans. Importantly, this increase was found only when the humans were depicted to be engaged in activities like playing cards or reading a book, underscoring the importance of perceived social behavior on eye movement control. However, only static pictures were used in these experiments, so that it remains unclear what would happen in more dynamic contexts typical for robot action.

An investigation of eye movement patterns in the context of dynamical stimuli was conducted by Klein and colleagues [37]. They examined mean fixation durations and counts while participants watched movies of two triangles. Some of the movements of the triangles gave rise to interpretations that made reference to the concept of "mind". Movies that were interpreted in terms of mind states, i.e., of one triangle "agent" exhibiting behavior that is related to the state of the other "agent", were associated with both longer fixation durations and a reduced number of fixations. Thus, this study yielded a first description of differential eye movement patterns between social and non-social situations. However, it did not provide a more detailed analysis of gaze control in social, as compared to non-social, situations. In particular, for modelling social gaze behavior, it is important to understand where humans direct their gaze. No such information was provided by [37]. Also, when used as an evaluative measure of social (as compared to non-social) interactions, fixation duration is likely to be too coarse a measure: it is strongly influenced by the movement velocities of the stimuli ("agents"), which may not provide a good cue to "agency". This problem is at least partly circumvented by additionally obtaining measures of fixation location, which are less influenced by the specific velocity profiles employed. Because of the importance of eye measures in social contexts, the current study was designed to extend the results of [37] by also analyzing the gaze positions relative to moving triangles and comparing them between movies for which mind attribution did or did not occur. This permitted testing whether social content really leads to more eye movements being directed to (social) agents. As this is a novel approach, data analysis is inevitably more exploratory than one would ideally wish.

2.2 Experiment

In the experiment, 12 healthy participants watched different movies, each displaying two triangles, a large (about 4°) red and a smaller (about 2°) blue one, that moved in a "self-propelled" manner for about 18 seconds. Earlier studies had shown that these movies could be grouped into either a random, goal-directed, or theory-of-mind condition [37, 38]; henceforth, these will be referred to by the terms non-social, goal, and social, respectively.[1] Each condition contained four movies. Movies in the goal condition typically lead to interpretations that involve interactions, for example, the small triangle is following the large triangle, or the two triangles move in a symmetric fashion. Social movies typically evoke mentalizing descriptions. An example would be that the large triangle is trying to motivate the small triangle to move out of the house, while the small triangle seems to be afraid of the outside. By contrast, movies from the non-social condition are described mostly in terms of physical movements, e.g., the triangles are floating around. See below for further descriptions of the animations.

In the present study, participants were asked to watch the movies attentively in order to report at the end what had happened. These reports were only used to check that participants had indeed concentrated on the movies (i.e., they were not further analyzed). Instead, the focus lay on the difference in gaze patterns among movies from the different conditions.

2.2.1 Materials and Procedure

Participants' eye movements while watching the movies were recorded using an SR-Research (Canada) Eyelink 1000 system (see Fig. 1). The sampling frequency of the eye tracker was 1000 Hz, and its accuracy is typically better than 0.5° of visual angle. All 12 participants had normal or corrected-to-normal vision and received either monetary compensation or course credits in exchange for their time. The movies they viewed were the Frith-Happé animations as used by [37, 38], shortened to 18 seconds each while preserving the essential story. The long version of the movies can be viewed under

[1] This terminology refers to the perception of the triangles as social agents, i.e., whether mind attribution to the triangles occurs or not.

Fig. 1 Experimental setup. Participants sat comfortably in a chair and watched the movies on a computer screen. Head stability was achieved by means of a chin rest.

`http://www.icn.ucl.ac.uk/dev_group/ufrith/research.htm#animations.`

As an illustration, Fig. 2 depicts snapshots from the social film "mocking".

Each participant saw all twelve movies in random order. After each movie, participants were asked to describe what they had seen.

2.2.2 Data analysis

All reported measures were first calculated for each movie and participant separately, prior to averaging across movies from the same condition.

Temporal Variables

Mean fixation duration was the average length of periods without blinks and saccades, that is, when the tracked eye was open, eye velocity was below $30\frac{^\circ}{s}$, acceleration below $8000\frac{^\circ}{s^2}$, and eye movement amplitudes were smaller than 0.5°. Similarly, *fixation counts* were calculated by counting the number of fixation periods that were interrupted by a blink, saccade event, or the end of the movie.

Spatial Variables

Trial time was defined as the length of the animation movies. *Triangle time* was the time eye gaze fell within a circle of 3° around the center of gravity of the red or blue triangle, respectively. *Top (mid) time* was the time eye gaze was within a circle of 1° around the top (middle) of the triangle (The top position was the edge with the longest sides.) The *percentage triangle time* was then given by

$$100 * \frac{triangle\ time}{trial\ time},$$

while *percentage top time* was calculated by

$$100 * \frac{top\ time}{top\ time + mid\ time}.$$

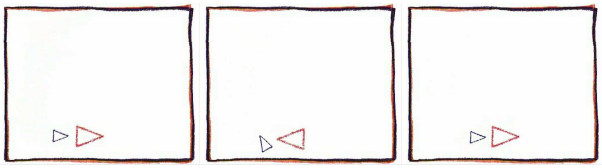

Fig. 2 Snapshots of the social film "mocking". A typical description for this film would be that the blue triangle mocks behind the back of the red one. When the red one turns and looks at the blue triangle the blue triangle pretends to do something else only to follow the red triangle afterwards again.

Percentage top time, however, was derived only for the red triangle; the small size of the blue triangle rendered this measure unreliable for this object. *Blue time* was the time eye gaze was within a circle of 3° around the center of gravity of the blue triangle, divided by triangle time. The *percentage middle time* was the time gaze was directed within 1° around the middle position between both triangles divided by trial time. Finally, a regression was calculated to examine whether the horizontal and vertical components of the eye movements were less dependent on the triangles' movements during social movies. For the regression, each movie was divided in four parts, each 4.5 seconds in length. Subsequently, for each part, a regression of the horizontal positions of the eyes x_e was run, using as regressors an intercept i, the horizontal position of the red triangle's center of mass x_r, the horizontal position of the blue triangle's center of mass x_b, and an error term ϵ; β_r and β_b are weighting parameters:

$$x_e = i + \beta_r x_r + \beta_b x_b + \epsilon$$

An analogous regression was run for the vertical position of the eyes y_e:

$$y_e = i + \beta_r y_r + \beta_b y_b + \epsilon$$

The crucial measure here was the proportion of accounted-for variance R^2. R^2 values were first averaged across all four parts and then across x_e and y_e to derive at the final measure of explained variance. For each measure, a repeated-measures ANOVA with the factor movie condition (non-social, goal, social) was calculated to examine for a general influence of the type of movie. Except for Blue time, this influence was followed up by planned comparisons between the goal and non-social and the social and goal conditions. As for Blue time, the percentage for each movie condition was tested against 50% using Bonferroni-corrected t-tests. Whenever necessary, Greenhouse-Geisser corrections [39] were used, though for ease of communication only the non-corrected degrees of freedom are reported below.

2.2.3 Results

Temporal Variables

Fig. 3 displays the mean differences in fixation duration and

fixation counts among the three movie conditions. Mean fixation durations were longer for movies in the social compared to the non-social and goal conditions, while the latter two conditions showed little difference in fixation duration. Similarly, fixation counts were highest in the non-social and lowest in the social condition, with the goal condition again being similar to the non-social condition. This pattern was reflected statistically in significant main effects of condition for fixation durations (F(2, 22) = 4.62, MSE = 1534.63, $p < .05$) and fixation counts (F(2, 22) = 4.29, MSE = 9.93, $p < .05$), and in significant differences between social and goal movies (fixation duration: F(1, 11) = 10.27, MSE = 1676.54, $p < .05$; fixation counts: F(1, 11) = 8.92, MSE = 13.72, $p < .05$), but no significant differences between non-social and goal movies (F(1, 11) = 1.38, $p > .10$; $F < 1$).

Spatial Variables

Fig. 4 - Fig. 8 show the results of the location analyses. Participants looked longer at the triangles in social movies compared to the other two conditions. In fact, this increase in looking time was nearly linear from non-social through goal to social movies (Fig. 4). Statistically, this was corroborated by a significant main effect of condition (F(2, 22) = 33.53, MSE = 76.14, $p < .05$; goal vs. non-social: F(1, 11) = 10.55, MSE = 106.71, $p < .05$; social vs. goal: F(1, 11) = 65.93, MSE = 56.76, $p < .05$).

A different pattern emerged for percentage top time. The social and goal movies exhibited a higher percentage of top time compared to the non-social condition, while differing only little between each other (Fig. 5). Accordingly, while there was a significant main effect of condition, only the comparison involving the non-social condition reached significance (F(2, 22) = 20.46, $p < .05$; goal vs. non-social: F(1, 11) = 38.53, MSE = 56.61, $p < .05$; social vs. goal: F(1, 11) = 2.48, MSE = 42.69, $p > .10$).

During movies of the non-social and goal conditions, gaze was more often directed to the blue rather than the red triangle. This ratio was more balanced for the social condition (see Fig. 6). Statistically, there was a main effect of condition, and Blue time differed significantly from 50% except in the social condition. (F(2, 22) = 13.57, MSE = 37.92, $p < .05$; t(11) = 3.93, $p_3 < .05$; t(11) = 6.98, $p_3 < .05$, t(11) = -.81, $p_3 > .10$ for the non-social, goal, and social conditions, respectively).

As can be seen from Figure 7, the time that was spent fixating midway between the two triangles increased significantly across the three conditions (F(2, 22) = 228.72, MSE =10.68, $p < .05$; goal vs. non-social: F(1, 11) = 352.78, MSE =9.30, $p < .05$; social vs. goal: F(1, 11) = 194.23, MSE =25.08, $p < .05$).

Finally, horizontal and vertical eye positions were significantly less determined by the locations of the triangles in social movies compared to the non-social and goal movies (F(2, 22) = 37.65, MSE <1, $p < .05$; goal vs. non-social: F(1, 11) < 1; social vs. goal: F(1, 11) = 47.56, MSE <0.01, $p < .05$). This difference in explained variance is shown in Figure 8.

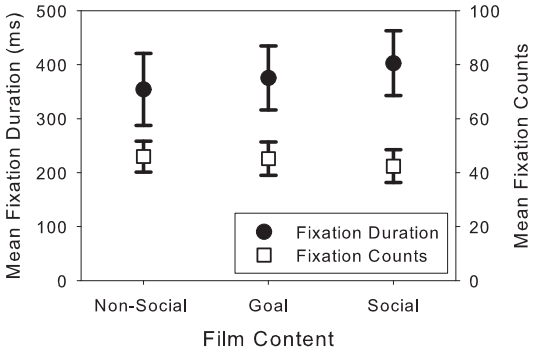

Fig. 3 Mean fixation durations and counts (and 95% confidence intervals) for social, goal, and non-social movies.

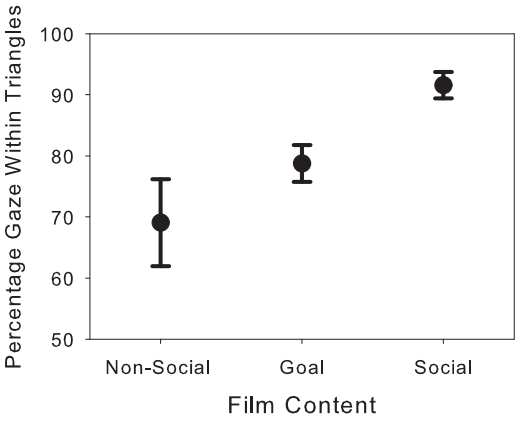

Fig. 4 Mean percentage trial time (and 95% confidence interval) for social, goal, and non-social movies.

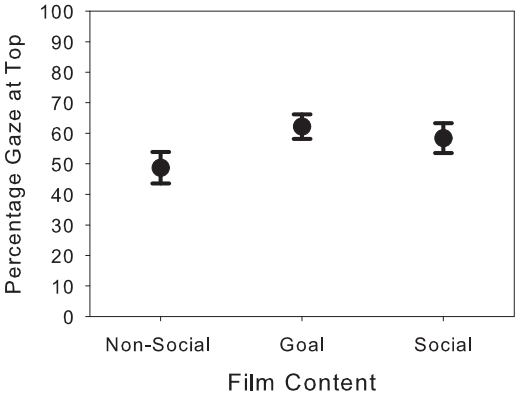

Fig. 5 Mean percentage top time (and 95% confidence interval) for social, goal, and non-social movies.

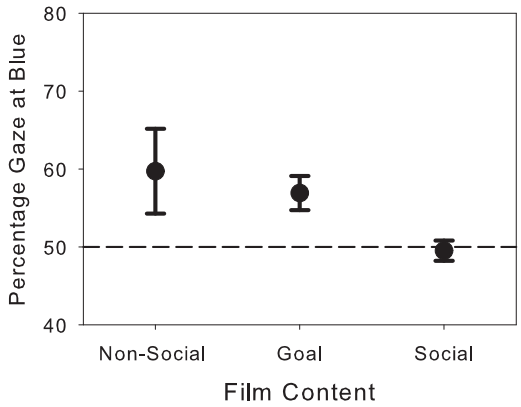

Fig. 6 Mean percentage blue time (and 95% confidence interval) for social, goal, and non-social movies.

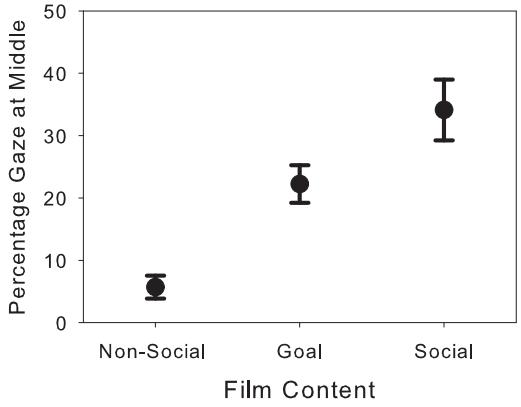

Fig. 7 Mean percentage time spent midway between both triangles (and 95% confidence interval) for social, goal, and non-social movies. Note that these values do not have to add up to 100% with those given in Figure 4 - as, due to the adopted size of the regions of interest, a fixation in-between the two triangles could in some extreme cases also be counted as fixating one of the triangles.

2.3 Discussion

In the experiment, participants watched films that showed two self-propelled moving triangles. As has been demonstrated before, these films typically lead to descriptions that mainly involve physical terms (non-social movies), refer to interactions (goal movies), or evoke mentalizing descriptions (social movies) [37,38]. While participants were watching these movies for subsequent report, their eye movements were recorded. Eye movements play a prominent role in social interactions (e.g., [40,4]) and are therefore well suited to investigating the activation of social interpretations during movies.

Indeed, fixations were significantly longer and less frequent while watching social movies. This fits well with the literature, where increases in fixation durations are reported when more complex processing [41,13,42], or integration

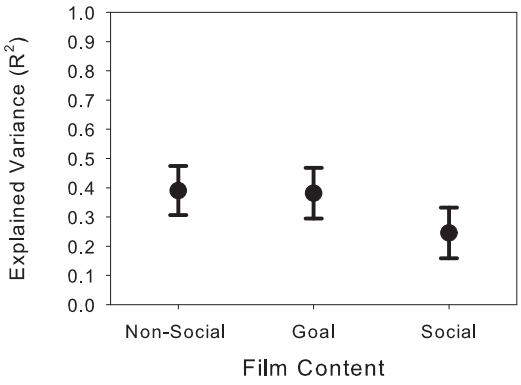

Fig. 8 Mean proportion of variance (and 95% confidence interval) accounted for by the triangle positions for social, goal, and non-social movies, combined across horizontal and vertical gaze positions.

of various types of information [43] is required. For example, [43] reported an increase in average fixation duration during cartoon viewing when the caption was provided in advance of, rather than after, the picture. The authors attributed this to the increased processing demands when integrating text and picture, which lead to longer average fixation durations. Clearly, social situations are situations where additional sources of information, besides the physical information, are to be integrated. This makes processing more complex in social conditions, thus increasing fixation durations.

In addition to this temporal variation, spatial differences were observed as well. The eyes remained longer on the triangles during social films and gaze position was more often placed at corner junctions of the red triangle which lay in its respective heading direction. Additionally, participants "dwelled" for more balanced amounts of the time on the red and blue triangles as well as the region between both. Further, gaze position was less dependent on the triangles' locations. Thus, these measures permit movies with mentalizing content to be reliably discriminated from those without such contents. Longer fixation durations on the triangles were expected on the assumption that watching social content activates a social model, which subsequently prioritizes eye movements to the "social agents", reflected in longer gaze times on the triangles. One might speculate that the longer fixation times on the corners corresponding to the triangles' heading directions in the social and goal conditions are related to a preference for looking at the eyes of intentional agents (e.g., [11]). Interestingly, this was the only measure in which the goal condition was more similar to the social than to the non-social condition. One reason for this might be that humans tend to look at the head of all entities that seem to show a movement that is goal-directed, whether or not the entity is attributed a mind. Goal-directed is meant here in the sense of having an end point that is designated

by the observable characteristics of the scene, rather than by a non-observable (mental) state. The more balanced distribution of gaze time over the red and blue triangles in the social condition shows that humans understand that, when both agents have a "mind" of their own, both agents have to be observed for a full understanding of the interaction.

This is less necessary in conditions where mind attribution does not occur, because it is sufficient to observe what one agent is doing to understand what is going on in such situations; here, goals are not seen as related to the state of the other agent. For example, understanding the blue triangle's action of "hiding" in a social condition requires that one understands at the same time that the red triangle is looking for someone. This is not true for the animation of "fighting" in the goal condition, where understanding the action of one triangle does not depend on the state of the other triangle. In these situations, visual factors, such as the (relative) size of the triangles, play a larger role, leading to a gaze preference for the harder-to-detect blue triangle.

The time gaze was directed to the middle position between the two triangles increased significantly from the non-social through the goal to the social condition. This suggests that the two triangles were perceived more as a common group [44] in the goal and social conditions, which made fixations to the individual triangles less necessary. [2]

Finally, the stronger influence of triangle location on gaze position during non-social movies suggests that theory-of-mind processes make eye control less dependent on visual input and physical events alone, and more dependent on internal models of what is happening and where to look next (see [34] for a similar argument in the context of goals). Importantly, except for top time, all measures differed between social and goal conditions - which demonstrates that they specifically index social processes.

An interesting question (raised by one anonymous reviewer) concerns at what point in time, after the start of the movies, the eye movement measures begin to dissociate between animations with social and those with non-social contents. Preliminary analyses of the data obtained with the current animations suggest that the reported measures differ in how fast they signal the detection and processing of social as compared to non-social material. While measures, such as middle time, triangle time, and blue time allow distinguishing between the conditions after about 4 seconds of sampling, it takes 14 seconds for top time to show a significant difference. This suggests that some measures more closely reflect relatively early, presumably visual processes that provide cues for social (as compared to non-social) contents, while others involve more time-consuming updating of the situational model and the activation of the appropriate oculomotor tracking routines. This possibility will have to be examined in future studies that systematically vary the visual cues and situational information cues rendered by the animations.

Importantly, these measures provide a quantitative index of how strongly a movie that is being watched activates social interpretations. This opens the possibility for using such measures to assess alternative designs for artifacts. For instance, it could answer questions such as: is it more important for evoking social interpretations to construct robot hands that are visually similar to human hands, or to move them with a biological velocity profile. One scenario, for example, could be that different prototypes of robots perform the same sequence of actions while human observers' eye movements are recorded. These eye movements can then be analyzed for the "markers" specified above to gain a measure of mind attribution. Similarly, these variables - the overall speed of fixation changes, and the time and location spent on intentional agents - should be taken into consideration when implementing algorithms for robotic camera control. This is important for providing humans with a more natural feeling when they interact with robots. Interesting in this context is the observation of [45], that humans are quite accurate in predicting the location of others' gaze. In future work, we plan to develop automatic classification routines for distinguishing between gaze patterns in social and non-social situations. These routines can then be used by a robot to select and evaluate its own gaze behavior. The aim will be to develop a robot that, in addition to establishing eye contact [46], is also capable of displaying human-like camera movements in the interaction that follows afterwards.

Having available more detailed knowledge about the eye movements of humans in social situations can inspire models of robot camera control, as well as inform how bottom-up-derived information should be complemented by top-down guidance. That is, in a given category of social situations, such as making first eye contact, camera control signals based on visual saliency could be overridden online by knowledge of where persons would typically look in such situations. In the example, this would permit eye contact to be maintained despite the presence of some salient (e.g., red) object on the background (e.g., green grass).

However, one caveat remains. While the current stimuli where selected based on the interpretations they evoke in the observer, these stimuli were less controlled in terms of their low-level movement characteristics. We attempted to counter this problem by using location information relative to the triangles, which should render the measure more robust against low-level feature differences than temporal measures alone. Additionally, we used 4 stimuli per condition. Optimally, physical differences should therefore be either representative of the category or cancel out. We contend that the displayed movements are a representative sample of naturally occurring movements of the same type, i.e., intentional agents do show sudden movement starts and stops.

[2] We thank one anonymous reviewer for this interesting suggestion.

However, further studies will have to show whether this assumption is valid or not.

The longer time needed for fixations in the social condition points to an interesting aspect only alluded to in the Introduction: perceiving a social actor to be present does not necessarily enhance the human's performance. Rather, as was the case, for instance, with humans observing incongruent movements in the study of Kilner and colleagues [20, 26], using social models to interpret others' actions can also be detrimental to performance. At the same time, however, having a good model of the other allows one to make more accurate predictions. Future challenges will therefore be to investigate more precisely what kind of (robot) implementations help improve human-robot interactions and exactly under which circumstances. We propose that eye tracking might be a useful technique to help find the right answers.

3 Conclusion

Eye movements were compared between movies that did or did not evoke mental-states attributions. We argued that these kinds of measures could also help to evaluate different implementations of robots. Future research will have to show where evoking mental interpretations can be beneficial (e.g., when predictions of others' behavior have to be correct) or detrimental (e.g., when simple repetitive actions have to be performed and interference should be minimized). Additionally, a deeper understanding of eye movement control will permit the development of robots better tailored for human-robot interaction.

Jan Zwickel received his Psychology and Ph.D. degrees from the Universities of Heidelberg and Leipzig, Germany. From 2007 until present, he has been working as a research fellow in the DFG-funded Excellence Cluster "Cognition for Technical Systems - CoTeSys" (see www.cotesys.org). His main research interests include human eye movement control in human-human as well as human-robot interaction scenarios, as well as interference effects in the performance of instructed motor actions by the concurrent perception of biological/non-biological movements. In 2002, he received, as a coauthor, the distinguished paper award of the IUI for his work on the evaluation of electronic tourist guides. Further, in 2009, he received the Otto-Hahn-Medaille of the Max-Planck-Society.

Hermann J. Müller received his Psychology degree from the University of Würzburg, Germany, and his PhD from the University of Durham, UK. Following a post-doctoral fellowship award by the German Research Foundation, he worked at the School of Psychology, Birkbeck College, University of London, UK. In 1997, he was appointed Chair of Experimental Psychology at the University of Leipzig. In 2000, he became Chair of General and Experimental Psychology at the Ludwig Maximilian University (LMU) Munich. He has a broad range of research interests including: visuo-spatial attention, adaptive weighting dynamics in visual search, cross-modal processing and motor action, and adaptive control and plasticity of cognitive functions. He uses a combination of behavioral, neuroscientific, and computational-modelling approaches. In 2007, he was awarded a special LMU Research Professorship, and in 2008 he was made a member of the LMU Center for Advanced Studies.

References

1. H. Deubel, W.X. Schneider, Vision Research **36**(12), 1827 (1996)
2. A.R. Hunt, A. von Mühlenen, A. Kingstone, Journal of Experimental Psychology: Human Perception and Performance **33**(2), 271 (2007). DOI 10.1037/0096-1523.33.2.271. URL http://dx.doi.org/10.1037/0096-1523.33.2.271
3. C. Breazeal, B. Scassellati., in *Proceedings of the 1999 IEEE/RSJ International Conference on Intelligent Robots and Systems* (1999), pp. 858–863
4. P. Persson, J. Laaksolahti, P. Lönnqvist, IEEE Transactions on Systems, Man, And Cybernetics - Part A: Systems and Humans **31**(3), 349 (2001)
5. D.H. Ballard, M.M. Hayhoe, P.K. Pook, R.P. Rao, Behavioural and Brain Science **20**(4), 723 (1997)
6. S. Johnson, V. Slaughter, S. Carey, Developmental Science **1**, 233 (1998)
7. Y. Nagai, K. Hosoda, A. Morita, M. Asada, Connection Science **15**, 211 (2003)
8. J. Triesch, C. Teuscher, G.O. Deák, E. Carlson, Dev Sci **9**(2), 125 (2006). DOI 10.1111/j.1467-7687.2006.00470.x. URL http://dx.doi.org/10.1111/j.1467-7687.2006.00470.x
9. K. Nation, S. Penny, Dev Psychopathol **20**(1), 79 (2008). DOI 10.1017/S0954579408000047. URL http://dx.doi.org/10.1017/S0954579408000047

10. S. Krach, F. Hegel, B. Wrede, G. Sagerer, F. Binkofski, T. Kircher, PLoS ONE **3**(7), e2597 (2008). DOI 10.1371/journal.pone.0002597. URL http://dx.doi.org/10.1371/journal.pone.0002597
11. E. Birmingham, W.F. Bischof, A. Kingstone, Visual Cognition **16**(2/3), 341 (2008)
12. S. Fletcher-Watson, J.M. Findlay, S.R. Leekam, V. Benson, Perception **37**(4), 571 (2008)
13. M.F. Land, S. Furneaux, Philosophical Transactions of the Royal Society London B **352**(1358), 1231 (1997). DOI 10.1098/rstb.1997.0105. URL http://dx.doi.org/10.1098/rstb.1997.0105
14. F. Heider, M. Simmel, The American Journal of Psychology **57**(2), 243 (1944)
15. B. Tversky, B.M. Hard, Cognition **110**(1), 124 (2009). DOI 10.1016/j.cognition.2008.10.008. URL http://dx.doi.org/10.1016/j.cognition.2008.10.008
16. G. Csibra, Cognition **107**(2), 705 (2008). DOI 10.1016/j.cognition.2007.08.001. URL http://dx.doi.org/10.1016/j.cognition.2007.08.001
17. J. Weizenbaum, in *Communications of the ACM*, vol. 9, ed. by A.G. Oettinger (1966), vol. 9, pp. 36–45
18. B. Duffy, Anthropomorphism and The Social Robot **42**(3-4), 177 (2003)
19. B. Hommel, J. Müsseler, G. Aschersleben, W. Prinz, Behavioral & Brain Sciences **24**(5), 849 (2001)
20. J.M. Kilner, Y. Paulignan, S.J. Blakemore, Current Biology **13**(6), 522 (2003)
21. G. Rizzolatti, L. Craighero, Annual Review of Neuroscience **27**, 169 (2004). DOI 10.1146/annurev.neuro.27.070203.144230. URL http://dx.doi.org/10.1146/annurev.neuro.27.070203.144230
22. V. Gallese, A. Goldman, Trends in Cognitive Sciences **2**(12), 493 (1998)
23. V. Gallese, Philos Trans R Soc Lond B Biol Sci **362**(1480), 659 (2007). DOI 10.1098/rstb.2006.2002. URL http://dx.doi.org/10.1098/rstb.2006.2002
24. T. Pozzo, C. Papaxanthis, J.L. Petit, N. Schweighofer, N. Stucchi, Behav Brain Res **169**(1), 75 (2006). DOI 10.1016/j.bbr.2005.12.005. URL http://dx.doi.org/10.1016/j.bbr.2005.12.005
25. L. Craighero, F. Bonetti, L. Massarenti, R. Canto, M.F. Destro, L. Fadiga, Brain Research Bulletin **75**(6), 770 (2008). DOI 10.1016/j.brainresbull.2008.01.014. URL http://dx.doi.org/10.1016/j.brainresbull.2008.01.014
26. J.M. Kilner, A.F. de C. Hamilton, S.J. Blakemore, Social Neuroscience **2**(3-4), 158 (2007). DOI 10.1080/17470910701428190
27. M. Grosjean, J. Zwickel, W. Prinz, Psychological Research/Psychologische Forschung **73**(1), 3 (2009). DOI 10.1007/s00426-008-0146-6. URL http://dx.doi.org/10.1007/s00426-008-0146-6
28. V. Gazzola, G. Rizzolatti, B. Wicker, C. Keysers, Neuroimage **35**(4), 1674 (2007). DOI 10.1016/j.neuroimage.2007.02.003. URL http://dx.doi.org/10.1016/j.neuroimage.2007.02.003
29. J. Stanley, E. Gowen, R.C. Miall, Journal of Experimental Psychology: Human Perception and Performance **33**(4), 915 (2007). DOI 10.1037/0096-1523.33.4.915. URL http://dx.doi.org/10.1037/0096-1523.33.4.915
30. F. Biocca, C. Harms, J.K. Burgoon, Presence **12**(5), 456 (2003)
31. D.C. Dennett, *The Intentional Stance* (Mit Press, 1989)
32. J.M. Zacks, Cognitive Science **28**(6), 979 (2004)
33. M. Hayhoe, D. Ballard, Trends in Cognitive Sciences **9**(4), 188 (2005). DOI 10.1016/j.tics.2005.02.009. URL http://dx.doi.org/10.1016/j.tics.2005.02.009
34. C.A. Rothkopf, D.H. Ballard, M.M. Hayhoe, Journal of Vision **7**(14), 1 (2007). DOI 10.1167/7.14.16. URL http://dx.doi.org/10.1167/7.14.16
35. K. Dautenhahn, I. Werry, in *Proceedings of the 2002 IEEE/RSJ Intl. Conference on Intelligent Robots and Systems* (2002), pp. 1132–1138
36. L. Brenner, K. Turner, R.A. Müller, Journal of Autism and Developmental Disorders **37**(7), 1289 (2007). DOI 10.1007/s10803-006-0277-9. URL http://dx.doi.org/10.1007/s10803-006-0277-9
37. A. Klein, J. Zwickel, W. Prinz, U. Frith, Q J Exp Psychol (Colchester) **62**(6), 1189 (2009). DOI 10.1080/17470210802384214. URL http://dx.doi.org/10.1080/17470210802384214
38. F. Abell, F. Happé, U. Frith, Cognitive Development **15**(1), 1 (2000). URL http://www.sciencedirect.com/science/article/B6W47-41GWN8R-1/1/732d0f8c8aa5ecca8adba01e4832d198
39. S. Greenhouse, S. Geisser, Psychometrika **24**(2), 95 (1959). URL http://dx.doi.org/10.1007/BF02289823
40. C. Breazeal, P. Fitzpatrick, in *Proceedings of the AAAI Fall Symposium Socially Intelligent Agents: The Human in the Loop* (2000)
41. D.E. Irwin, *The interface of language, vision, and action* (New York: Psychology Press, 2004), chap. Fixation location and fixation duration as indices of cognitive processing
42. G. Underwood, L. Jebbett, K. Roberts, The Quarterly Journal of Experimental Psychology A **57**(1), 165 (2004). DOI 10.1080/02724980343000189. URL http://dx.doi.org/10.1080/02724980343000189
43. P.J. Carroll, J.R. Young, M.S. Guertin, *Eye movements and visual cognition. Scene perception and reading* (NewYork: Springer-Verlag, 1992), chap. Visual analysis of cartoons: A view from the Far Side, pp. 444–461
44. H.M. Fehd, A.E. Seiffert, Cognition **108**(1), 201 (2008). DOI 10.1016/j.cognition.2007.11.008. URL http://dx.doi.org/10.1016/j.cognition.2007.11.008
45. S.W. Bock, P. Dicke, P. Thier, Vision Research **48**(7), 946 (2008). DOI 10.1016/j.visres.2008.01.011. URL http://dx.doi.org/10.1016/j.visres.2008.01.011
46. D. Miyauchi, A. Sakurai, A. Nakamura, Y. Kuno, in *CHI '04: CHI '04 extended abstracts on Human factors in computing systems* (ACM, New York, NY, USA, 2004), pp. 1099–1102. DOI http://doi.acm.org/10.1145/985921.985998