2019 Joint IEEE 9th International Conference on Development
and Learning and Epigenetic Robotics (ICDL-EpiRob)
Oslo, Norway, August 19-22, 2019

New evidence for learning-based accounts of gaze following: Testing a robotic prediction

Priya Silverstein
Psychology Department
Lancaster University
Lancaster, UK
p.silverstein@lancaster.ac.uk

Gert Westermann
Psychology Department
Lancaster University
Lancaster, UK
g.westermann@lancaster.ac.uk

Eugenio Parise
Psychology Department
Lancaster University
Lancaster, UK
e.parise@lancaster.ac.uk

Katherine Twomey
Division of Human Communication, Development & Hearing
University of Manchester
Manchester, UK
katherine.twomey@manchester.ac.uk

Abstract——**Gaze following is an early-emerging skill in infancy argued to be fundamental to joint attention and later language. However, how gaze following emerges has been a topic of great debate. The most widely-accepted developmental theories suggest that infants are able to gaze follow only by understanding shared attention. Another group of theories suggests that infants may learn to follow gaze based on low-level social reinforcement. Nagai et al. [Advanced Robotics, 20, 10 (2006)] successfully taught a robot to gaze follow purely through social reinforcement, and found that the robot learned to follow gaze in the horizontal plane before it learned to follow gaze in the vertical plane. In the current study, we tested whether 12-month-old infants were also better at gaze following in the horizontal than the vertical plane. This prediction does not follow from the predominant developmental theories, which have no reason to assume differences between infants' ability to follow gaze in the two planes. We found that infants had higher accuracy when following gaze in the horizontal than the vertical plane (*p* = .01). These results confirm a core prediction of the robot model, suggesting that children may also learn to gaze follow through reinforcement learning. This study was pre-registered, and all data, code, and materials are openly available on the Open Science Framework (https://osf.io/fqp8z/).**

Keywords—cognitive development, developmental robotics, gaze following, reinforcement learning

I. Introduction

The ability to engage in joint attention (the shared focus of two individuals on an object) is a key developmental milestone, and an important precursor to language development [1]. However, how joint attention initially emerges is the subject of great debate. Arguably a critical precursor to joint attention is gaze following: orienting to the same object that another interlocutor is looking at [2], [3]. Gaze following is found in human infants (e.g. [4], [5]) and many non-human species (e.g. [6], [7], [8], [9]). Evidence for the age at which this ability initially emerges is mixed, however. Whether or not infants are able to gaze follow in a given study seems to be affected not only by age but also whether stimulus presentation is live or computerized, whether there is perceived movement, whether a change in eye gaze direction is accompanied by a head turn, whether a communicative or attention directing cue is used before the gaze shift, and what measure is used (for a review, see [10]). Eye-tracking studies of gaze following show some evidence of gaze following from 5 months of age [11], [12], [13], [14], including in non-WEIRD (Western, educated, industrialized, rich, and democratic) populations [15].

While gaze following in infancy has been repeatedly demonstrated, the origins of this skill remain controversial. Some influential developmental theories assume that it is an innate ability which relies on infants' capacity to understand others' communicative intent (e.g. [16], [17]). Specifically, on these accounts infants are born with the understanding that adults are intentional communicative agents, and based on this understanding, look where adults look in order to obtain that information. In support of these accounts are studies that suggest infants may have some rudimentary form of gaze following (gaze cueing) from birth, evidenced by the finding that newborns detect an object on the screen faster if it appears in a location previously cued by another's gaze ([4]). Others argue that although this is not an ability that is present from birth, at a certain time point the ability to read others' intentions and hence follow gaze 'switches on'. This account is supported by the finding that younger infants need the presence of ostensive cues (direct gaze and infant directed speech) in order to follow gaze ([13], but see [12], [14] for the argument that this is due only to the attention grabbing qualities of these cues). More broadly, while overall there exists a family of rich accounts of gaze following, all incorporate one or more innate, presumably genetically encoded components assumed to exist from birth or to come online according to some maturational timetable.

In contrast to these rich interpretations, recent computational work has proposed a range of lower-level processes which may support the development of gaze following without the need for any understanding of intention. Rather, these simulations offer mechanisms by which infants may learn this skill, in particular through social reinforcement [18], [3], [19]. Specifically, if adults tend to look at things that are interesting, then infants are positively reinforced when they look in the same direction, even if initially this is just by chance. Modeling work thus demonstrates that the typical development of gaze following can be simulated with a combination of the infant's perceptual skills and preferences, habituation and reward-driven learning, and a structured social environment in which the caregiver tends to interact with objects that the infant finds interesting. In particular, Nagai, Asada and Hosoda [20] successfully taught a developmental robot with a simple neural network cognitive architecture to gaze follow through low-level, supervised associative learning, without any built-in understanding of intentionality. The robot was equipped with a camera which fed images of a human experimenter to a visual system consisting of a connectionist map, which encoded these images, and a retinal smoothing layer, simulating the development of infants' visual acuity. In the experimental set-up the experimenter held up an object in the robot's visual field. After processing the visual input from the camera, the robot generated a motor command, adjusting the joint angles in its head and neck, resulting in a head turn and a change in its visual field. The robot was then given feedback based on the output error between the location of the object in the visual field and its gaze direction: if the object was centered in a predetermined location in the visual field, gaze following was considered successful and no adjustments to the neural network were made. When the object was outside the visual field or off-center, gaze following was considered unsuccessful, and random noise was added to the connection weights in the robot's neural network. Across training, therefore, head movements resulting in incorrect gaze following were less likely to be produced, increasing the relative strength of connections which produced correct gaze following[1].

Following training, testing with previously untrained images demonstrated that the robot could successfully follow the experimenter's gaze. However, testing at intervals in training revealed that it did so in stages: early in learning, the robot initially learned to follow gaze in the horizontal plane, and only later in the vertical plane. Importantly, horizontal input to the robot was more perceptually variable than the vertical input, suggesting that in this model, environmental input was critical in shaping the trajectory of the robot's behavioral development. Thus, this work makes the empirically testable prediction that infants too should initially follow gaze shifts more successfully in the horizontal than the vertical direction. However, potential differences in infants' ability to follow gaze in the horizontal and vertical plane are yet to be explored. More broadly, this system learned to follow gaze based on visuomotor input and associative learning, raising the possibility that low-level perceptual and proprioceptive information coupled with social reinforcement are sufficient to support the emergence of this important ability.

In this study, we test the prediction made by Nagai and colleagues' [20] robotic model. Specifically, we ask whether 12-month-old infants are better at gaze following in the horizontal than the vertical plane. If infants show better gaze following in the horizontal than the vertical plane, this would confirm a central prediction of the model of gaze following that is based on low-level mechanisms. Our piloting showed that 12-month-old infants were able to follow the gaze direction in our stimuli, but were not at ceiling or floor. This age group are therefore at an intermediate developmental stage in which gaze following is not consistently accurate, raising the possibility that at this age, we may observe the differences in horizontal/vertical tracking predicted by the robot. All pre-registered hypotheses, materials, code, and data can be found on the Open Science Framework (OSF; https://osf.io/fqp8z/).

II. Method

A. Participants

16 typically developing 12-month-old infants took part in the experiment (mean age: 364 days; range: 352 days to 376 days; 11 female; all Caucasian; 13 monolingual English). One additional infant was excluded due to fussiness. All infants were reported to have no developmental delays and no visual impairments that would stop them being able to see the stimuli.

B. Stimuli and Design

Example stimuli are depicted in Figure 1. Each trial consisted of a three-second long video. Trials were split into control trials and experimental trials. Control trials consisted of a central fixation cross (1000 ms) followed by a novel object appearing in one of four locations 200 pixels left, right, up, or down from the center of the screen (2000 ms), and were designed to test whether infants found gaze shifts (without gaze following) *a priori* easier in the horizontal than the vertical plane. Experimental trials consisted of a human face looking directly at the infant, surrounded by four images of the same object in each of the locations (1000 ms) followed by averted gaze to one of the four locations (2000 ms). We used five photographs of one female face looking left, right, up, down, and directly at the camera. The eyes from the left, right, up and down photographs were superimposed onto the face looking directly at the camera, ensuring the face was identical apart from gaze direction. We selected objects from the NOUN database [21] that had no obvious top/bottom in order to avoid biasing infants' attention. In total, 64 videos were made, consisting of the eight objects in the four locations for both the control and experimental conditions. These videos were pseudorandomized into four orders such that infants never saw the same object, location or trial type (control, experimental) on more than two successive trials. All videos are available on OSF.

C. Procedure

Infants sat on their caregiver's lap during the experiment in front of a 23-inch screen (seated approximately 0.6 meters away). An eye-tracker (Tobii X120) captured infant looking times and gaze locations on screen. We used Tobii Studio 3.3.1 to present stimuli and gather eye-tracking data. We performed a five-point calibration for all infants before beginning the experiment. After this calibration, we instructed caregivers not to talk to or interact with their infant, and that they could stop

[1] A formal description of the model is outside the scope of the current paper; for details see Nagai et al. (2006)

the experiment at any time if the infant became too fussy, and the experiment began. Infants saw up to a maximum of 64 trials in one of the four pseudorandomized orders. Caregivers decided when to end the experiment, with 12 infants seeing the full 64 trials.

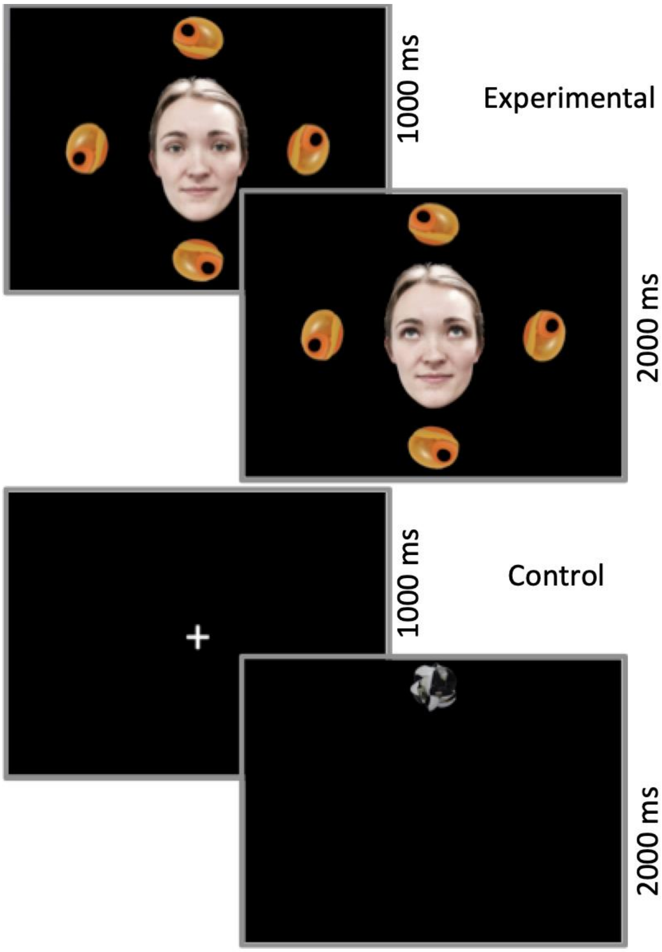

Fig. 1. Experimental and control conditions. Central cue (direct gaze or fixation cross) for 1000 milliseconds, followed by a directional cue (averted gaze or object) for 2000 milliseconds.

D. *Analysis*

We performed all analyses in R 3.5.2 [22]. Analyses by plane were preregistered on OSF. After visualizing the data and observing differences between the positions within the planes, we performed complementary exploratory analyses by position. We exported raw data from Tobii Studio 3.3.1 and analyzed them using the eyetrackingR package [23]. First, Areas Of Interest (AOIs) were defined for center (400 x 400 pixels), and left, right, up, and down object locations (each 340 x 400 pixels). Our accuracy variable was proportion looking: length of looking at target object divided by total looking at all four objects.

III. Results

Table 1 provides infants' average looking times to each possible stimulus location following object appearance (control trials) and gaze shifts (gaze trials).

TABLE I. Average looking time (ms) to all locations for all trial types (control, gaze). Congruent locations are indicated in bold.

Trial	*Loc^n*	**Average Looking Time (ms) Post Onset of Shift/Object**				
		Centre	***Down***	***Left***	***Right***	***Up***
Control	Down	402	**1130**	6	3	40
Control	Left	331	40	**1167**	13	23
Control	Right	371	24	9	**1127**	33
Control	Up	440	62	10	6	**1099**
Gaze	Down	794	**186**	274	290	106
Gaze	Left	896	137	**213**	270	125
Gaze	Right	820	215	231	**300**	104
Gaze	Up	832	148	258	275	**109**

A. *By Plane*

We submitted proportion looking to a 2 x 2 repeated measures ANOVA with main effects of condition (gaze vs. control) and plane (horizontal vs. vertical) and their interaction, with planned follow up paired *t*-tests (Figure 2). The ANOVA revealed a significant interaction between condition and plane for proportion looking [$F(1,15) = 10.28$, $p = .006$]. There was a significant main effect of condition, with accuracy being significantly higher for control than gaze trials [$F(1,15) = 959$, $p < 0.001$]. There was a significant main effect of plane, with accuracy being significantly higher for the horizontal than the vertical plane [$F(1,15) = 6.59$, $p = .02$]. We carried out separate paired *t*-tests to assess the effect of plane on each condition. In gaze trials, accuracy was significantly higher for the horizontal [$M = 0.33$, $SD = 0.16$] than the vertical plane [$M = 0.17$, $SD = 0.11$; $t(15) = 2.81$, $p = .01$]. In control trials, there was no significant difference in accuracy across the two planes [Horizontal: $M = 0.91$, $SD = 0.09$; Vertical: $M = 0.93$, $SD = 0.06$; $t(15) = -1.16$, $p = .27$].

B. *By Location*

We submitted proportion looking to a 2 x 4 repeated measures ANOVA with main effects of condition (gaze vs. control) and position (up, down, left, and right) and their interaction, with planned follow up one way repeated measures ANOVAs and paired *t*-tests (Figure 3). The ANOVA revealed a significant interaction between condition and position for proportion looking [$F(3,45) = 4.78$, $p = .006$], a significant main effect of condition (as above), and a significant main effect of position [$F(3,45) = 4.67$, $p = .006$]. We carried out one way repeated measures ANOVAs to assess the effect of position on each condition. In gaze trials, there was a significant effect of position [$F(3,45) = 5.26$, $p = .003$], whereas in the control trials, there was not [$F(3,45) = 0.61$, $p = .61$]. We carried out separate paired *t*-tests to assess the effect of position in the gaze trials. In gaze trials, accuracy was significantly lower for the up looks [$M = 0.09$, $SD = 0.13$] than the down looks [$M = 0.24$, $SD = 0.22$; $t(31) = 2.35$, $p = .03$], left looks [$M = 0.29$, $SD = 0.24$; $t(31) = 2.32$, $p = .03$], and right looks [$M = 0.36$, $SD = 0.15$;

t(31) = 3.42, *p* = .002]. There was no significant difference in accuracy between down and left looks [*t*(31) = -0.11, *p* = .91], down and right looks [*t*(31) = -1.42, *p* = .17] or left and right looks [*t*(31) = -1.01, *p* = .32].

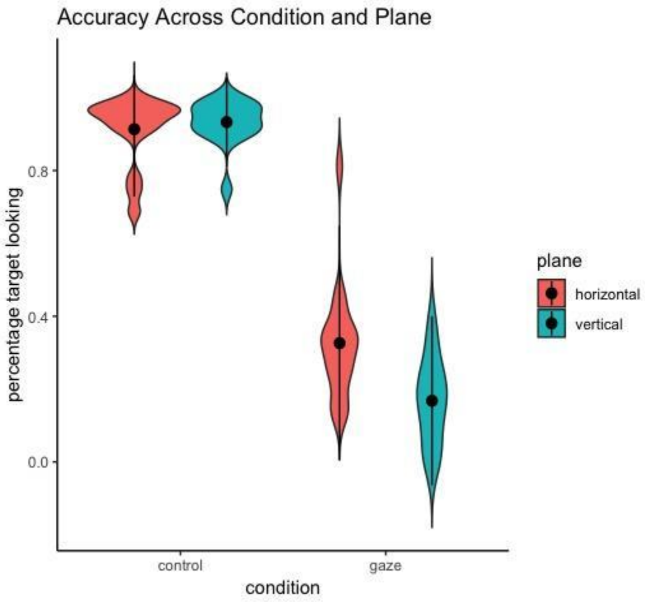

Fig. 2. Gaze following accuracy across condition and plane. Point range shows means and standard deviations. Width shows probability density.

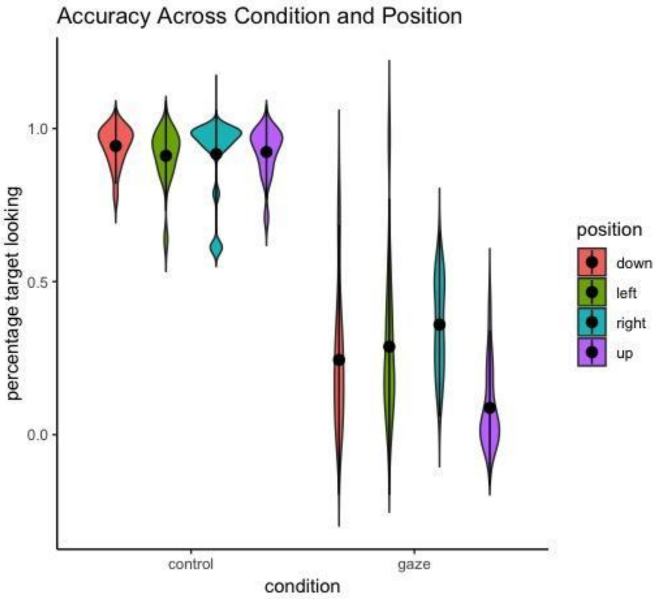

Fig. 3. Gaze following accuracy across condition and position. Point range shows means and standard deviations. Width shows probability density.

Discussion

In the current study we tested the predictions of Nagai et al's [20] robotic implementation of the development of gaze following. In this work, the robot learned to gaze follow based on associative learning and visuomotor input, and did so more rapidly in the horizontal than the vertical plane. Our results capture this prediction: they suggest that 12-month-old infants have higher accuracy when gaze following in the horizontal than the vertical plane. We find no such difference in our control condition, where objects appear in the same locations but are not cued with gaze. Importantly, the results of our control condition therefore indicate that the difference found in the experimental condition is not due to *a priori* more accurate eye movements in the horizontal than vertical plane. Our results are therefore compatible with the robotic prediction, in which gaze following emerged from a combination of associative learning mechanisms and social reinforcement.

It is interesting to note that infants in our study are 'bad' at gaze following, that is, they show a maximum of 36% accuracy in the right look condition (where 25% is chance performance), and spend most of their time fixating the center of the face (see Table 1). This is to be expected, as it has been shown repeatedly that infants respond best to head and eye movements in combination [24]. The movement of just the eyes is a very subtle cue, and lacks head movement information, and so is understandably more difficult for infants to follow. We chose to manipulate only eyes in this experiment in order to create stimuli that were as controlled as possible, but future research should explore whether our findings can be replicated for combined head movement and eye cues, since if gaze following is learned, these additional cues should also be integrated by the learning mechanism.

Our interpretation rests on the assumption that in the real world, infants encounter more variability in gaze input in the horizontal plane than the vertical plane. Indeed, in Nagai and colleagues' input to the robot, greater horizontal variability in the gaze input interacted with the robot's developing visual acuity leading to more successful initial learning in the horizontal plane. A further test of this mechanism would be to increase vertical variability in the input to the robot; in this case, if the robot learned to gaze follow first in the vertical plane, this would support this explanation of the model's behavior. Importantly, however, whether infants too encounter more horizontal than vertical variability in gaze information is not known, pointing to an interesting empirical question for future work. These two types of evidence in combination with the current study would provide strong support for the theory that infants can learn to follow gaze based on the interaction of input, low-level associative processes and social reinforcement

While the current data offer support for learning-based accounts of gaze following, we cannot rule out alternative accounts based on innate cognitive processes. In particular, the shape of the human eye is such that horizontal eye movements are easier to see, due to the amount of visible white sclera being larger. For this reason, one could argue for an innate perceptual system tuned to spot horizontal eye movements (and see [4], for evidence of gaze cueing in newborns). However, if this were the case, in the current study we might expect to find better gaze following in the upwards vertical gaze (where there is more visible sclera) than the downwards vertical gaze (where no sclera is visible), which is not the case (see Figure 3). In fact, interestingly, we find significantly worse accuracy for upwards looks than all three other positions. Intuitively this is consistent with learning theories, as infants are typically situated below adults' line of gaze, which would elicit more frequent downwards than upwards looks from the adult. Again, however, empirical evidence from naturalistic input to infants is needed to assess this possibility. Nonetheless, since in this study we are not studying learning itself but rather an outcome of the possible learning mechanism, we cannot rule out the possibility that some other innate mechanism biases gaze following to horizontal planes. However, we are not aware of any such theory making this prediction. Furthermore, our control condition goes some way towards controlling for this, as if information in the horizontal direction is generally richer, we might expect differences also when objects appear in horizontal and vertical locations without being cued by gaze. One way in which a developmental account could be further strengthened is by testing if the observed horizontal-vertical difference is also present in infants at a lower age.

Taken together, our results and those of Nagai et al. [20] raise the possibility that a low-level learning mechanism provided with structured input may be sufficient to support the development of gaze following in human infants; whether infants do additionally possess an innate ability to read others' intentions remains an open question. Importantly, however, current prominent theories in developmental psychology (e.g. [16], [17]) do not predict that there would be differences in accuracy of gaze following across the horizontal and vertical planes, as an understanding of shared attention or communicative intent (the basis for these theories) is not dependent on direction of gaze. As such, the current results point to new opportunities for the development of mechanistic theories that can account for these behavioral data. More broadly, the current study highlights the important contribution developmental robotics can make to the testing and refinement of developmental theory.

Acknowledgment

We would like to thank Malcolm Wong for help with data collection, Katharina Kaduk for her role as lab manager at Lancaster Babylab, and the infants and caregivers who took part in the study.

References

[1] M. Tomasello, "Joint attention as social cognition," in *Joint attention: Its origins and role in development*, Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc., 1995, pp. 103–130.

[2] R. Brooks and A. N. Meltzoff, "The development of gaze following and its relation to language," *Dev. Sci.*, vol. 8, no. 6, pp. 535–543, Nov. 2005.

[3] C. Moore, M. Angelopoulos, and P. Bennett, "The role of movement in the development of joint visual attention," *Infant Behav. Dev.*, vol. 20, no. 1, pp. 83–92, Jan. 1997.

[4] T. Farroni, S. Massaccesi, D. Pividori, and M. H. Johnson, "Gaze Following in Newborns," *Infancy*, vol. 5, no. 1, pp. 39–60, 2004.

[5] M. Scaife and J. S. Bruner, "The capacity for joint visual attention in the infant.," *Nature*, vol. 253, no. 5489, pp. 265–266, 1975.

[6] T. Bugnyar, M. Stöwe, and B. Heinrich, "Ravens, Corvus corax, follow gaze direction of humans around obstacles," *Proc. R. Soc. Lond. B Biol. Sci.*, vol. 271, no. 1546, pp. 1331–1336, Jul. 2004.

[7] S. Okamoto-Barth, J. Call, and M. Tomasello, "Great Apes' Understanding of Other Individuals' Line of Sight," *Psychol. Sci.*, vol. 18, no. 5, pp. 462–468, May 2007.

[8] D. J. Povinelli, T. J. Eddy, R. P. Hobson, and M. Tomasello, "What Young Chimpanzees Know about Seeing," *Monogr. Soc. Res. Child Dev.*, vol. 61, no. 3, p. i, 1996.

[9] 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.

[10] A. Frischen, A. P. Bayliss, and S. P. Tipper, "Gaze cueing of attention: Visual attention, social cognition, and individual differences.," *Psychol. Bull.*, vol. 133, no. 4, pp. 694–724, Jul. 2007.

[11] G. Gredebäck, C. Theuring, P. Hauf, and B. Kenward, "The Microstructure of Infants' Gaze as They View Adult Shifts in Overt Attention," *Infancy*, vol. 13, no. 5, pp. 533–543, Sep. 2008.

[12] G. Gredebäck, K. Astor, and C. Fawcett, "Gaze Following Is Not Dependent on Ostensive Cues: A Critical Test of Natural Pedagogy," *Child Dev.*, vol. 89, no. 6, pp. 2091–2098, Nov. 2018.

[13] A. Senju and G. Csibra, "Gaze Following in Human Infants Depends on Communicative Signals," *Curr. Biol.*, vol. 18, no. 9, pp. 668–671, May 2008.

[14] J. Szufnarowska, K. J. Rohlfing, C. Fawcett, and G. Gredebäck, "Is ostension any more than attention?," *Sci. Rep.*, vol. 4, no. 1, May 2015.

[15] M. Hernik and T. Broesch, "Infant gaze following depends on communicative signals: An eye-tracking study of 5- to 7-month-olds in Vanuatu," *Dev. Sci.*, p. e12779, Dec. 2018.

[16] S. Baron-Cohen, "Mindblindness An Essay on Autism and Theory of Mind," p. 6.

[17] G. Csibra and G. Gergely, "Natural pedagogy," *Trends Cogn. Sci.*, vol. 13, no. 4, pp. 148–153, Apr. 2009.

[18] G. O. Deák, A. M. Krasno, J. Triesch, J. Lewis, and L. Sepeta, "Watch the hands: infants can learn to follow gaze by seeing adults manipulate objects," *Dev. Sci.*, vol. 17, no. 2, pp. 270–281, Mar. 2014.

[19] J. Triesch, C. Teuscher, G. O. Deak, and E. Carlson, "Gaze following: why (not) learn it?," *Dev. Sci.*, vol. 9, no. 2, pp. 125–147, Mar. 2006.

[20] Y. Nagai, M. Asada, and K. Hosoda, "Learning for joint attention helped by functional development," *Adv. Robot.*, vol. 20, no. 10, pp. 1165–1181, Jan. 2006.

[21] J. S. Horst and M. C. Hout, "The Novel Object and Unusual Name (NOUN) Database: A collection of novel images for use in experimental research," *Behav. Res. Methods*, vol. 48, no. 4, pp. 1393–1409, Dec. 2016.

[22] R Core Team, *R*. 2017.

[23] J. W. Dink and B. Ferguson, *eyetrackingR: An R Library for Eye-tracking Data Analysis*. 2015.

[24] C. Moore, M. Angelopoulos, and P. Bennett, "The role of movement in the development of joint visual attention," *Infant Behavior and Development*, VOL. 20, no. 1, pp. 83-92, Jan 1997