One-year old infants control bottom-up saliencies to purposely sustain attention

Andrés H Méndez[13], Chen Yu[2], Linda B Smith[1]

amendez86@gmail.com
chen.yu@austin.utexas.edu
Smith4@indiana.edu
[1]Department of Psychological and Brain Sciences, Indiana University
[2]Department of Psychology, University of Texas at Austin
[3]CICEA, Universidad de la República

Salient stimuli attract gaze [1,2]. Mature perceivers internally suppress salient distractors to purposefully sustain attention on a visual target. Infants' abilities to purposefully sustain gaze on an object, often measured in the context of play, is also assumed to require the internal suppression of distractors and is considered an early marker and risk point in the development of the internal regulatory processes mediated by the pre-frontal cortex [3,4]. Here we show that sustained attention by one-year-old infants includes a behavior-driven increase in the external salience of the target. Using head-mounted eye trackers, we measured infants' gaze during object play and the momentary visual size of objects in the infant's field of view. Visual size is well-known to robustly attract gaze [1]. We found that when infants directed gaze to an object, there was a simultaneous change in the the spatial relation of the head to the attended object increasing the target's visual size relative to distractors. The onset, duration, and offset of the increased salience was time-locked with the onset, duration and offset of infant gaze to the object. The findings challenge characterizations of infant attention as a competition between bottom-up and top-down control and implicate instead a collaboration in which top-down goals drive infant's externally-directed behaviors that suppress the salience of distractors at input. The top-down control of attention through externally directed behavior may serve as the training ground –and risk factor – in the development of internal control.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Participants

Parents of all participating infants gave written informed consent prior to participation. Recruitment and all procedures were approved by the Human Subjects Review Board of Indiana University (IU protocol number 0808000094). Dyads were recruited from outreach events and an opt-in data base from Monroe County, Indiana. The final sample consisted of 45 parent-infant dyads (25 male, mean age 14.1 months, SD = 1.4). Fifteen additional dyads began the study but did not contribute but were not included because they did not tolerate the sensors, were fussy, or equipment problems.

Setup

Infants and parents sat across from each other at a small table (61cm × 91cm × 64cm). Infants sat on a chair in front of their parents. The table, floor and walls were white and participants wore white smocks and the toys in play were each a different uniform color which aided in the algorithmic coding of images for the toys and their image size by pixel color. The infant wore head-mounted eye trackers from Positive Science, LLC. Tracking system included an infrared camera – mounted on the head and pointed to the right eye of the participant – that recorded eye images, and a scene camera that captured the events from the infant's perspective. The scene camera's visual field is 108 degrees, providing a broad view to approximate the full visual field Both the scene and eye camera recorded at 30Hz. Three additional cameras recorded the interaction from third-person views.

Stimuli

Each child played with two different sets of three novel toys each with volume of about 300 cm^3. The toys were designed and constructed of various materials and moveable parts by the experimenters to be engaging to 1 year old infants. Within each set of 3 toys, one toy was a uniform color blue, green and red. To increase generalizability of the findings, different infants played with different sets of 3 toys. In total,

across the 45 infants, 18 different unique toys were used.

Procedure

The head gear was placed and adjusted on the infant as an experimenter and parent kept the infant occupied and in active play with a toy. Fifteen calibration points at different points on the play table were collected; the experimenter directed the infant's attention toward an attractive toy used only for calibration while the technician recorded the attended moment that was used in later eye tracking calibration. Because infants this age do not explore or play actively with objects on their own, or with strangers, parents assisted in the experimental procedure. They were told that the goal of the experiment was to study how infants explored and interacted with objects during play and were asked to interact with their child and keep them engaged with the toys as naturally as possible. Each infant received their two unique sets of toys each containing 3 objects. Infants played with each set twice in four 1.5 minute trials in the order of 1212, with the sets randomly assigned as set 1 or 2. Because young infants often did not actively explore or look at the toys for the full 1.5 minutes, we selected the first 45 sec of each trial for analysis, a period during which all included infants were actively looking at the toys.

Data processing and coding

The quality of eye-tracking video for each dyad was checked to ensure calibration quality. Each infant contributed 5400 frames (45 sec * 4 trials * 30Hz), a total corpus of 243,000 frames. Analyses were conducted on images in which gaze was directed to at least one of the toys in the infants scene-camera image and the looked at object was in view. Of the 158.296 frames with a look to an object in only 2951 frames was the looked-at object not in view (1.9%). As a result, the corpus analyzed includes 155046 on-task frames (63,8% of the total corpus).

Scene properties

The presence, location and image size of toy objects was algorithmically determined from the infant's head camera raw images. This was accomplished using computer vision techniques in three different steps. The first step separates object and background pixels by the color of pixels in the image that were unique to each object and on white background. The second step groups non-white pixels into larger blobs to determine which portions of the image belong to the same object even if that object is broken up visually into multiple segments, as for example, when held in a participant's hand. The third step assigns each blob into an object category. Once the number of pixels belonging to each object was determined for all frames, the validity of the automatic coding results were assessed by asking two human coders to annotate a small proportion of the data (~ 1200 frames); the comparison of these hand codings with the image processing results yielded 91% frame-by-frame agreement.

For each image frame (sampled at 30Hz), we determined the visual size (VS, degrees subtended), and the relative visual size (RS, proportion of the object pixels in the image that belonged to each object) of each object in the image. RS was calculated directly from the number of image pixels belonging to each of the objects in view compared to the sum of all the pixels belonging to any object in that frame:

$$RS_1 = \frac{\#\ pixels_1}{\sum_{i=1}^{n} \#\ pixels_i}$$

being RS_1 the relative size of an object, $\#pixels_1$ the number of image pixels belonging to that object and n the number of objects in view.

The visual size of each object on a frame was calculated from the degrees in a single pixel and the number of pixels belonging to the corresponding object. Infant's head camera videos are 480 pixels in height by 640 pixels in width, which at 72 dpi, translates to 22.cm x 16.9 cm, respectively. The average distance of the eye to the table center for infants sitting on the chair was 44.5 cm. The degrees in a single pixel of the head camera is then:

$$\theta_{pix} = \frac{\arctan\left(0.5 \times 16.9 \big/ 44.5\right)}{0.5 \times 480}$$

To calculate the visual size we assumed that each object projected as a circle to the camera. The visual size for each object was calculated from the object's projected diameter and θ_{pix}.

Gaze and looking behavior

Frame by frame gaze was determined for three regions-of-interest (ROIs), the three toy objects in each scene-camera image. Highly trained coders, naive to the hypotheses and goals of this study, indicated when the gaze cross-hair fell on a pixel belonging to any of the three ROIs. Because the three toys in play were three different primary colors different from skin tones and the white background, this is done with high accuracy. A second coder independently coded a randomly selected 10% of the frames with the inter-coder reliability ranged from 82% to 95% (Cohen's kappa = 0.81). A look to an object was defined as an unbroken continuous stream of frames in which gaze was within the same object ROI. A sustained look was defined as an unbroken look that was 3 secs or longer.

Salience stability during a look

Potential looks were defined as unbroken gaze within the region of single ROI (the pixels belonging to one object. Two looks were combined into one look if broken by no more than one frame (.33 sec). For each frame within a look, it was determined whether the object that was looked at was also the absolute largest object in that frame (at least .34 of all object pixels). The stability of the RS (relative size) advantage was computed as the proportion of frames within a look for which the looked to object was the largest. The RS baseline was computed as the mean of a random selection of one object in each frame (independent of gaze) in the corpus and thus reflects the null hypothesis that looking is random with respect to RS.

Quantification and statistical analyses

Mixed-effect linear regression models were conducted at the corpus level using the lme4 package in R (Version 3.6.1; Bates, Mächler, Bolker, & Walker, 2014). RS was the fixed effect and intercepts were specified for individual infants and for the specific target objects.

RESULTS

During free play, the *visual* size of objects varies continuously with body and object movements that alter the proximity, view, or partial occlusion of objects in the perceiver's field of view. For each image frame (sampled at 30Hz), we determined the visual size (VS, degrees subtended) and the relative visual size (RS, proportion of the object pixels in the image that belonged to each object) of each object in the image. The VS and RS of an object within a frame were positively correlated (**Figure 1B**, r^2= 0.52). We used frame-by-frame RS as the principal measure of the relative salience of competing visual targets. In virtually all images (>.99), one object was visually larger than the other two objects and typically was considerably so. If all three objects are in view and the same visual size, RS is expected to be .33 for all objects; the mean RS of the largest object (varying by definition from .34 to 1.00) was 0.63 (.17) and 76% had a RS greater than .50 indicating that one object was larger than the sum of the distractors (**Figure 1C**).

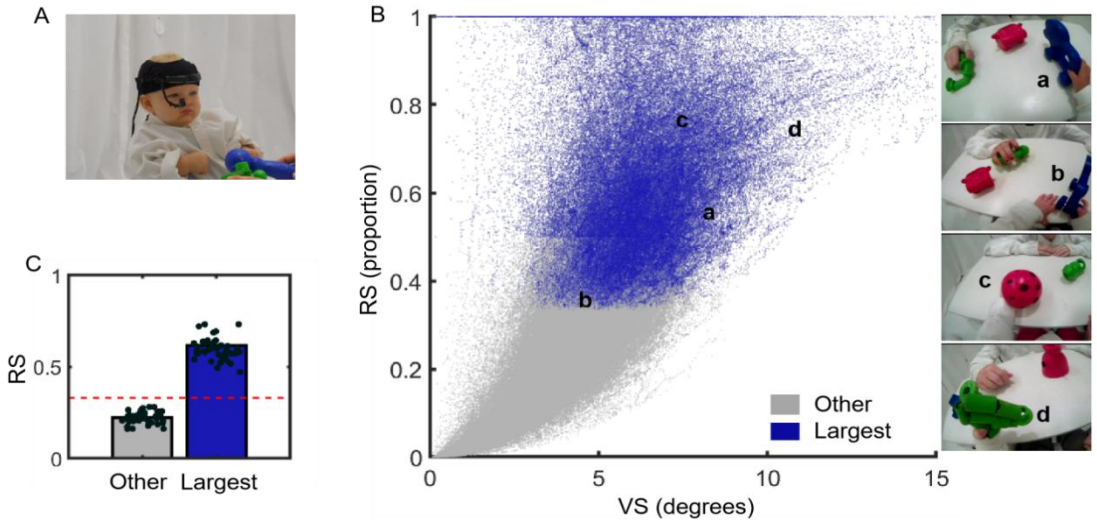

Figure 1. ***A:*** Infants wore head-mounted eye trackers while playing with 3 toys of the same physical size. ***B:*** The frame-by-frame visual size (VS) and relative visual size (RS) of the 3 toys in play and example head camera showing for examples of objects with the largest visual size in the image. ***C:*** The bar graph show the corpus means of RS for the largest object and the other objects in each frame; the dot clouds show the corresponding participant means. The dotted red line indicates the expected value of RS if all 3 toys were the same visual size. The authors received signed consent for the infant's photograph to be published in this article.

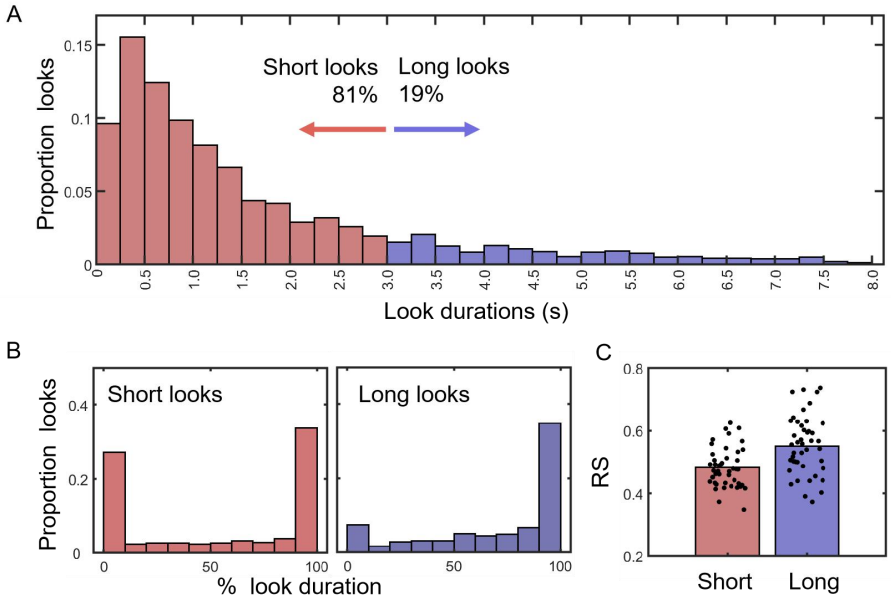

Figure 2. *A:* The distribution of look duration (bin size 250 msec.). Looks equal to or longer than three seconds (purple) are categorized as Long looks. ***B.*** The distribution of Short and Long looks in which the looked-to object is the largest in the infant's field of view. ***C***. Bar height indicates corpus mean of RS for the duration of Short and Long looks and the cloud of dots show the participant means.

We measured look durations as continuous gaze to a single object. Infant look durations are known to be extremely skewed [5], with most lasting less than a second but with a long tail of long looks (**Figure 2A**). Long looks, associated with sustained attention, predict the later development of self-regulation and executive function [e.g., 4, 6, 7]. Following established approaches [8, 9, 10], we used a threshold of look durations greater than or equal to 3 seconds to define Short and Long looks. Frame-by-frame RS could in principle vary substantially within a single look to the same object, if the distance of object to the head varies during the look. We determined the stability of frame-by-frame RS for the duration of each look using as our index the proportion of each look duration for which the attended object was the largest object in the image. The resulting distributions of durations (**Figure 2B**) were different (Kolmogorov Smirnov, p <<0.001) for Short (<3 sec) and Long looks (≥ 3 sec). For Short looks (<3 sec), the distribution was bimodal: the attended object either remained the visually largest or remained *not* the visually largest for the near entirety of a look to an object. For Long looks, the distribution was unimodal; the attended object was the largest in the infant field of view for the near entirety of the look. The mean RS of the attended object during long looks (M_{long} = 0.55; SE_{long} = 0.0097) was reliably greater than the mean RS of the attended object during Short Looks (M_{short} = 0.48; SE_{short} = 0.0048; β = 1.34; z = 5.72; p < 0.001, **Figure 2C**). Thus, the RS for both Short and Long looks were categorically stable, remaining either the visually largest or not the visually largest for the duration of the look. However, it was specifically Long Looks that were strongly associated with a target visually larger than competitors for the duration of the look.

During Long Looks, the salience advantage of the attended object was time locked to the onset and offset of the look (**Figure 3**). To determine when the RS of the attended object diverged from a baseline RS calculated across all in-view objects, we determined the first significant difference from baseline in a series of ordered pairwise t-tests from 2 secs before the onset and the last significance difference (return to baseline) from 2 sec before offset [11]. By these measures, the RS of the attended object reliably increased at 100 msec before the look onset and reliably decreased at 200

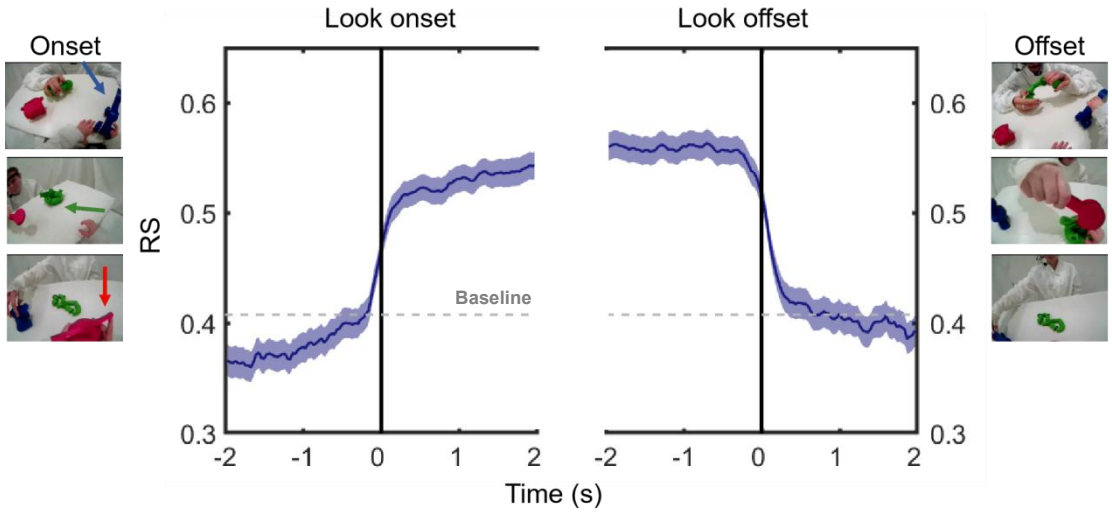

Figure 3. Relative size at onset and offset of Long looks. Mean RS of the looked-to object and standard error from 2 secs before to 2 secs after onset and offset of the look. The grey dotted line shows the baseline of RS for a sample of randomly selected objects. The head camera images show three examples of the RS of the looked-to object (indicated by arrows in the onset images) at the frame coinciding with coded onset and offset

msec. after the offset. Thus, the onset and offset of the salience advantage of the target during Long looks was nearly simultaneous with the onset and offset of the look. .

The observed jump in visual size at the onset of the look and its decrease at offset can arise from multiple events during play, including the infant's movement of the head, or by hand actions (by the infant or the parent). These movements can increase visual size by decreasing the distance of the attended object to the infant's head and eyes or by removing occlusions from other objects. Analyses of parent and infant handling (**Figure 4**) strongly indicate that hand actions on the objects did not play the critical role in creating the salience advantage. Although handling was common (98% of all Long Looks included some handling, 13% by parent, 25% by infant, and 60% by both parent and infant) these events were not stable across a sustained look,and were not temporally aligned with onset and offset of a look. These findings implicate changes in head position as principally

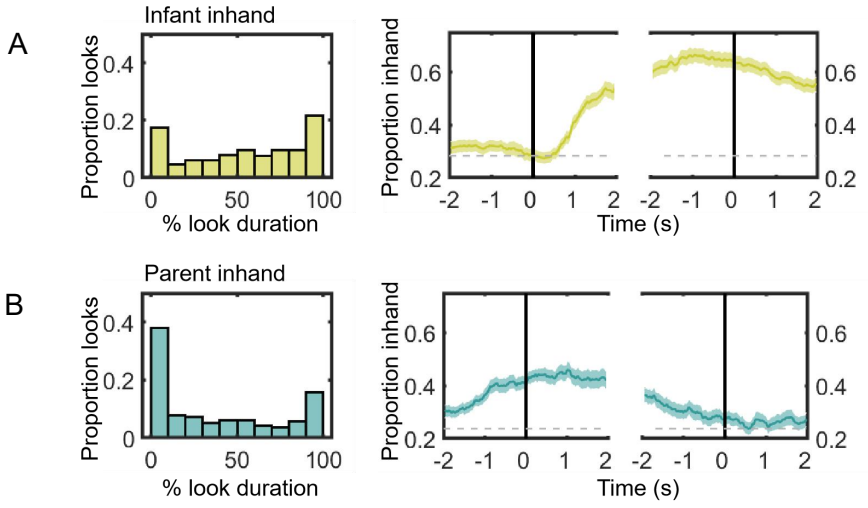

Figure 4. ***A.*** The histogram of the durations of long looks in which the looked to object was in the infants hands shows that infant handling of the looked to object during long looks was neither stable nor continuous. The onset and offset of infant handling was not temporally coordinated with the onset and offset of the look. The first significant difference [11] in infant handling from a baseline (determined as the mean of infant handling a randomly selected object across all frames) was 733 msec after onset. Infant handling did not reliably return to baseline within the +/- 2 sec window around look offset. ***B***. The histogram of the durations of Long looks in which the infant looked-to object was in the parent hands also was not stable during Long looks and not temporally coordinated with the onset and offset of the look. Parent handling was above baseline (determined a the mean of parent handling a randomly selected object across all frames) and did not change in the +/- window around look onset. The first significant difference [11] in parent handling from a baseline was 2233 msec before onset (before 2 second window) and a return to baseline at 766 msec after offset.

responsible for the change in RS at onset and offset of the look. The timing of the observed changes in RS at offset and onset coincides with what is known about the timing of head and eye movements during gaze shifts in freely moving infants, [e.g., 12-14]. Although coordinated head and eye movements at the start and finish of infant looks to a target has been documented [12], the role of these coordinated movements in creating a salience advantage during sustained attention is a new contribution. Given the synchronicity of the salience advantage with the onset of looking, the function of coordinated head and eye movements appears to not be in attracting attention to targets of interest but in helping the infant suppress distractors once that target is chosen on other top-down grounds.

DISCUSSION

Many biological organisms move their sensory surfaces toward a target of interest to control attention and to support the extraction of information [e.g., 15-17]. In primates, eye movements also support information extraction by selectively increasing the visibility of the target over competitors as the retinal area around the gaze point captures a higher resolution image than does the periphery. For targets of strong interest (as indicated by Long Looks), infants in the present study placed an additional gain on the visibility of a selected target by moving their body so that the target was visually larger than the competitors for the duration of the look. In this way, infants' top-down interest in an object used externally directed behavior to mitigate the need for the internal suppression of distractions to sustain attention.

Growing theory and evidence suggests that infants' experiences in sustaining attention plays a causal role in the development of executive control processes [3, 18, 19]. The present findings suggest that these executive functions may begin by controlling externally directed behaviors that resolve competition by controlling the input [20]. A mechanistic path from external to internal control of competition may reside in the known overlap of brain networks that plan eye-head movements and those that control visual attention [21] and in the feed-forward and feedback connections between pre-frontal cortex and early stage visual processing evident in infancy [3,18]. Identifying the early experiences that support the development of executive functions is a major goal of developmental science given the strong evidence that individual differences form early, are experience dependent, and have life-long consequences [3, 22]. The present findings open new directions of study on how infant externally directed behaviors in the service of controlling sustained visual attention may recruit and support the development of the brain's attention networks.

ACKNOWLEDGMENTS

The research was funded by NIH grants R01 HD074601 and R21 EY017843 and NSF BCS 1842817.

REFERENCES

[1] Borji, A., Sihite, D. N., & Itti, L. (2013). What stands out in a scene? A study of human explicit saliency judgment. Vision research, 91, 62-77.

[2] van Renswoude, D. R., Visser, I., Raijmakers, M. E., Tsang, T., & Johnson, S. P. (2019). Real‐world scene perception in infants: What factors guide attention allocation?. Infancy, 24(5), 693-717.

[3] Rosen, M. L., Amso, D., & McLaughlin, K. A. (2019). The role of the visual association cortex in scaffolding prefrontal cortex development: A novel mechanism linking socioeconomic status and executive function. Developmental cognitive neuroscience, 39, 100699.

[4] Brandes-Aitken, A., Braren, S., Swingler, M., Voegtline, K., & Blair, C. (2019). Sustained attention in infancy: A foundation for the development of multiple aspects of self-regulation for children in poverty. Journal of experimental child psychology, 184, 192-209.

[5] Suarez-Rivera, C., Smith, L. B., & Yu, C. (2019). Multimodal parent behaviors within joint attention support sustained attention in

infants. Developmental psychology, 55(1), 96.

[6] Johansson, M., Marciszko, C., Gredebäck, G., Nyström, P., & Bohlin, G. (2015). Sustained attention in infancy as a longitudinal predictor of self-regulatory functions. Infant Behavior and Development, 41, 1-11.

[7] H.A. Ruff, K.R. Lawson, R. Parrinello, R. W eissberg. Long-term stability of individual differences in sustained attention in the early years. Child Development, 61 (1990), pp. 60-75

[8] Yu, C., Suanda, S. H., & Smith, L. B. (2019). Infant sustained attention but not joint attention to objects at 9 months predicts vocabulary at 12 and 15 months. Developmental science, 22(1), e12735.

[9] Yu, C., & Smith, L. B. (2016). The social origins of sustained attention in one-year-old human infants. Current biology, 26(9), 1235-1240.

[10] Wass, S. V., Clackson, K., Georgieva, S. D., Brightman, L., Nutbrown, R., & Leong, V. (2018). Infants' visual sustained attention is higher during joint play than solo play: is this due to increased endogenous attention control or exogenous stimulus capture?. Developmental science, 21(6), e12667.

[11] Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. Journal of memory and language, 38(4), 419-439.

[12] Borjon J.I.*, Abney D.H., Yu C., Smith L.B., (In Press). Head and eyes: Looking behavior in 12- to 24-month old infants. Journal of Vision.

[13] Schmitow, C., Stenberg, G., Billard, A., & Hofsten, C. V. (2013). Using a head-mounted camera to infer attention direction. International Journal of Behavioral Development, 37(5), 468-474.

[14] Kretch, K. S., & Adolph, K. E. (2015). Active vision in passive locomotion: Real‐world free viewing in infants and adults. Developmental Science, 18(5), 736-750.

[15] Kleinfeld, D., Ahissar, E., & Diamond, M. E. (2006). Active sensation: insights from the rodent vibrissa sensorimotor system. Current opinion in neurobiology, 16(4), 435-444.

[16] Hofmann, V., Sanguinetti-Scheck, J. I., Künzel, S., Geurten, B., Gómez-Sena, L., & Engelmann, J. (2013). Sensory flow shaped by active sensing: sensorimotor strategies in electric fish. Journal of Experimental Biology, 216(13), 2487-2500.

[17] Taub, M., & Yovel, Y. (2020). Segregating signal from noise through movement in echolocating bats. Scientific Reports, 10(1), 1-10.

[18] Ellis, C. T., Skalaban, L. J., Yates, T. S., & Turk-Browne, N. B. (2021). Attention recruits frontal cortex in human infants. Proceedings of the National Academy of Sciences, 118(12).

[19] Katsuki, F., & Constantinidis, C. (2012). Early involvement of prefrontal cortex in visual bottom-up attention. Nature neuroscience, 15(8), 1160-1166.

[20] Byrge, L., Sporns, O., & Smith, L. B. (2014). Developmental process emerges from extended brain–body–behavior networks. Trends in cognitive sciences, 18(8), 395-403.

[21] Fiebelkorn, I. C., & Kastner, S. (2020). Functional specialization in the attention network. Annual review of psychology, 71, 221-249

[22] Hackman, D. A., Gallop, R., Evans, G. W., & Farah, M. J. (2015). Socioeconomic status and executive function: Developmental trajectories and mediation. Developmental science, 18(5), 686-702.