

Joint Visual Attention

A Tiny Filament Connecting the Disciplines of Cognitive Science

Ian Fasel

ianfasel@cogsci.ucsd.edu

December 10, 2000

1 Introduction

In the last decade, forty years since the beginning of the “cognitive revolution”, many cognitive scientists have been experiencing a paradigm shift, in which emphasis has moved from the representations that exist within a single individual to the cognitive phenomena that take place *between* individuals. The elaboration of these social skills has even been described as the primary adaptation of our species (Johnson, 1999; Hutchins & Strum, 1995); we have become experts in coordinating our activities with each other, and interacting in mutually beneficial ways. There are many different levels at which one may study this issue. On a higher level, we wish to understand how our behaviors are constrained by our social interactions and how we develop and use the skills needed to participate in these social interactions. On the lower level, we wish understand exactly what the physical basis for these interactions is – what kinds of data are our perceptual systems actually providing to us, and how are our brains putting those pieces together in a way optimally suited to allowing us to have the kinds of complex social behaviors we do.

In this paper, I examine the topic of joint visual attention, a behavior which is

foundational to many of the more complex social interactions of human beings, but which also occurs to varying degrees in other animals, and which lends itself to both a biological and computational account. In doing so, I hope to explore the more general question of what exactly will be required for a complete account of the coordinated activity of multiple human beings, using as a guide what might be described as the principle metaphor of cognitive science – that of cognition as computation (Hutchins, 1995; Marr, 1982).

2 Shared Attention in Development

Joint (or shared) visual attention is often thought of as a principal, foundational skill on which social coordination with others is built. In its simplest form, joint visual attention (or ‘deictic gaze’) is defined simply as ‘looking where someone else is looking’ (Butterworth, 1991, p. 223). This behavior is considered to be the first step towards the ability to share experience with others and to negotiate shared meanings. This in turn allows an infant to leverage the knowledge of the adult in order to learn about their environment, in part by manipulating the attention of the adult and in part by forming the basis for more complex forms of communication such as gestures and language.

When the problems of development are stated in Piagetian terms, joint attention provides a much-needed mechanism for explaining both assimilation and accommodation as they occur in stages of a child’s development. In the assimilation period of a stage, deictic gaze is an explanation of how particular aspects of the world are selected for attention and assimilation into a schema. If a child (for some presumably biological reason) feels compelled to attend to whatever the caretaker is attending to, the result may be a more optimal allocation of resources for learning (since, among all the things in the world that could be attended to in a day, the things the caretaker attends to are more likely to be salient with respect to the problems encountered in daily life). Meanwhile, in

the accommodation phase, once a particular schema has reached a critical level of disequilibrium, a child's ability to share attention may facilitate the ability to discover and take on the solutions to the problems encountered which are available in the society. These solutions, stored in the minds of the child's caretakers and social group, constitute the culture which the child is becoming a part of.

Thinking in these terms has led some to suggest that culture can be thought of as an apparatus for storing solutions to commonly encountered problems (Cole, 1985; Hutchins & Strum, 1995). While in humans these solutions are often stored in physical artifacts, they are also stored in social arrangements and strategies held within individual minds. Because knowledge is distributed and shared, it is then not necessary for each individual to independently develop the complexity of the entire society alone. Instead, problems are solved through a long process of development in which new skills are incrementally bootstrapped from previously acquired skills, many of which come "ready made" in the society. By this account, the child in many cases is often not constructing new skills to solve common problems, but rather is coming into coordination with preexisting skills stored in society (Cole, 1985; Johnson, 1999; Vygotsky, 1978; Hutchins & Strum, 1995).

Descriptions of shared attention, whether visual or in another form (such as auditory or gestural), is thus a specific, detailed account of the process of development, a level of description that Piaget's theories have been criticized for leaving out. Because it links the outside, social world to the internal, perceptual mechanisms of the brain, it is one of the few domains currently available for which a coherent story can be told connecting the neurological processes of perception and action to the social behaviors observed in everyday life.

3 The Development of Deictic Gaze

The development of shared visual attention follows a rapid and stereotyped sequence which has been fairly closely studied by developmental psychologists. One of the earliest results, reported by Scaife and Bruner (1975), reported that infants could adjust their gaze in response to a change in focus of their mother as young as two months of age. A number of other studies (see Scheffer, 1984 for a review) showed that mothers also closely monitor the gaze of their infants from very early on, meaning that shared attention occurs in both directions from almost the first moments of an infant's life in which she is alert and awake.

A number of studies conducted by Butterworth and colleagues, carried out under strictly controlled conditions, were able delineate three stages for the development of joint visual attention in children aged six to 18 months (for reviews see Butterworth, 1991 and Butterworth, 1995). In the first stage, children at age six months seated facing their mothers were able to determine from their mother's gaze what side of the room to look for interesting objects, but were not able to zero in on exactly which object was being attended and were not able to look behind them. In this case, children observing their mother fixate on a target in front of them would turn in the correct direction and then fixate on the first object of interest that appeared along the scan path. If the mother's gaze was directed at an object behind them, the infant would either fixate on an object in front of them or would not respond.

Butterworth referred to this as the 'ecological' mechanism, noting that according to Grover (1988), adding movement to the correct target could increase the accuracy of nine month olds to 100 percent. Butterworth hypothesized that infants at this age are able to use the gaze of other's as a cue for where to search for interesting objects, but it is the intrinsic interest value of the objects (often the same interest value that initially attracted the mother's attention) that completes the communicative function of the mother's head turn.

At twelve months, infants are better able to localize the gaze of their mother and show a new behavior of fixating intently on her during the turn, pausing an average of one second while she is still, and then turning their head rapidly to gaze at the target. Butterworth called this the ‘geometric’ mechanism, hypothesizing that infants were now computing an invisible line between mother and referent, and thus were able to pass over interesting objects in the scan path in order to fixate the correct target. However, infants at this age are still unable to access the space behind them, suggesting that they are not yet using a representation of being in a space.

The ability to look behind them (provided that there are no interesting targets in front of them) does not come until eighteen months of age, when infants are able to turn around to look at fixation points behind them if no interesting objects are in sight in front of them. This third stage is labeled the ‘representational’ stage, because Butterworth (1991) hypothesizes that the infant is making use of a represented space containing the infant as well as other objects outside the immediate visual field.

4 Mechanisms

Butterworth and colleagues’ account reports a set of fairly reliable data as well as a developmental theory. While many experimenters have confirmed the generic course of the onset of behaviors associated with gaze following (with some variation, e.g., Deak, Flom, & Pick, 2000 showed that even 12-month-olds will look behind them under certain conditions), the developmental theory posed by Butterworth and colleagues is somewhat speculative due to lack of data. Rather than developing a new representational mechanism, perhaps joint attention is learned through reinforcement, after months of interaction involving changes in gestures and facial expressions (Corkum & Moore, 1995; Deak et al., 2000). Furthermore, Butterworth and colleagues provide little data that help to pick apart

what specific parts of the interaction between child and caregiver are important in eliciting a gaze following response, thereby making it difficult to determine if the changes in behavior are due to changes in internal representations or simply changes in perceptual skills.

In order to answer these questions, several studies have been done varying different parameters of gaze-following tasks in the hope that specific mechanisms might be revealed. For instance, Corkum and Moore (1998) were able to demonstrate that 8-9 month-olds who had not previously shown spontaneous gaze-following could be trained to do so, giving credence to the theory that gaze following is a learned behavior. However, in the same study, Corkum and Moore also showed that 8-9 month-olds not showing spontaneous gaze-following behavior could not be trained to look in a direction *opposite* the direction of the experimenter's gaze, implying that simple learning alone is not sufficient as the mechanism through which joint attention is acquired.¹

Alternatively, many researchers have investigated the perceptual features (or, 'signal releasers', as described by Johnson, Slaughter, & Carey, 2000), which may provide specific cues for children to follow gaze without having to have any theory of intentionality about the caretaker. Moore, Angelopoulos, and Bennett (1997) found that while 9 and 12 month-olds were able to use just the head turn movement to correctly orient towards a target, the final static head orientation alone was only sufficient for previously spontaneous gaze-followers, and was not sufficient for training 9-month-olds. The same study also found that replacing the head-turn with a head-tilt, in which the experimenter essentially pointed towards a target with the crown of her head while keeping her eyes fixed on the

¹However, just because the infants were not eliciting gaze-following at 8-9 months does not mean that this skill was not in some as-yet-unrecognizable stage of development. Thus, while the correct condition might have been enough to 'push them over the edge' developmentally, so to speak, the contrary condition may not have been enough to undo all the previous learning that had taken place. Thus, the conclusion that this experiment proves that simple learning is not sufficient is not convincing, in my opinion.

infant, was not sufficient for any of the babies, although there was some weak evidence that spontaneous 12-month olds could learn this form of pointing. The implication of these studies is that some features of the gazer in the act of gazing are more important than others, especially the features of motion and faces.

In his master's thesis, Movellan (1986) tested the theory that it is not the specific features of human faces that are important, but rather the fact that one side of our head (our face) moves and makes sounds contingently on the behavior of others, especially the infant it is interacting with. To test this theory, 12-month-olds were seated on their mothers' laps in front of a remote controlled, non-humanoid, cube-shaped robot that had a non-facelike array of lights on one surface. During the experiment, the decorated side of the robot was oriented towards the baby and the lights were flashed and sounds were emitted in response to the child's vocalizations. In the experimental condition, similar looking robots were placed at the sides of the room and the center robot would occasionally interrupt its interaction with the infant in order to orient towards and exchange beeps and light flashes with the side robots. The entire sequence of exchanges between the remote-controlled robots and the infant were recorded. A control condition using a different infant was then done identically to the experimental condition, except that rather than the robot behaving contingently to the child, a previous recording of an interaction was used, so that the robot was behaving identically as in a contingent experiment except it was no longer behaving contingently to the child in front of it.

There were two results of this experiment. First, it was clear that the infants were extremely engaged and excited by the interactions with the contingent robots, but very quickly became bored with the non-contingent robot. Second, in the experimental condition, a significant number of babies learned to orient towards the side robots during the inter-robot interactions, suggesting that they were following the 'gaze' of the primary robot. A similar experiment, inspired by this one and conducted by Johnson et al. (2000) over a decade later, found

again that oriented but faceless creatures could elicit gaze-following from 12-month-olds if they behaved contingently, but not if their behavior was non-contingent but instead produced by a recording (identical to the Movellan (1986) experiment). This study also included another condition in which the object had a face on one end but did not behave contingently. In this case, gaze following behavior was still elicited. In the combination case, in which a non-living object with a face behaved contingently, there was a slightly higher incidence of gaze-following than in either no-face contingent or the face not-contingent conditions.

5 Features of Shared Attention

There are many possible interpretations of the results presented in the previous section. One possibility is that increased ability to follow gaze over time does not represent a change in the representational skills of infants, but rather a change in the perceptual skills of infants. This is supported by experiments that varied angle of head-turn, size of gesture, and complexity/interest value of target (Moore et al., 1997; Deak et al., 2000), finding that twelve-month olds were less likely to differentiate or even respond to more subtle stimuli than older children.

These studies also suggest that infants may use very specific features in gaze following. In particular, they suggest that oriented motion is necessary but not sufficient, that both facial features and contingent activity are sufficient but not necessary, and that the combination of any of these is sufficient to produce gaze-following. This raises questions about the kinds of perceptual skills needed to detect these features, the mechanism behind development of these feature detectors, and the mechanism linking the detection of these features to the behavior of gaze following.

The ability to detect ‘eyes looking at me’ is a skill that is quite prevalent

among vertebrates, and is documented even in the hognosed snake, chickens, lizards, the blue crab, ducks, and other primates (Baron-Cohen, 1995, suggests Ardouino & Gould, 1985 for a review). This has led to the suggestion that an innate ability to detect self-directed eyes exists in most vertebrates (Baron-Cohen, 1995), and that in primates, who are the only species capable of gaze following in cases other than ‘eyes looking at me’, this evolutionary adaptation has been especially aided by the development of a lighter sclera (the white of the eyes in humans), supposedly making it easier to determine eye direction (Johnson et al., 2000).

These arguments are certainly plausible; being aware of whether a predator has fixated its gaze on oneself is certainly of import to any animal that is preyed upon, and the fact of the change in the color of the sclera in the human eye is a very compelling fact. However these theories are difficult to test behaviorally, and there is little behavioral evidence to help explore the precise nature of these questions. Thus from this point on it seems necessary to appeal to levels of description that can explain the mechanisms underlying the mechanisms of shared attention (which has itself been described as a mechanism underlying social interactions such as gestures and language!) It is for this reason that I now turn to a model described by Baron-Cohen, and an attempt to implement that model in a humanoid robot.

6 The Baron-Cohen Model

In an attempt to make a biologically and computationally plausible account of how shared visual attention is produced, Baron-Cohen (1995) distinguished between dyadic and triadic relations in order to give a symbolic description of the gaze-following problem and a Fodorian model of how the problem is solved.

Baron-Cohen’s model describes dyadic representations as those following the form [*Agent-Relation-Agent*], in which *Self* or *Proposition* can fill the *Agent* slot.

Examples of these are [*Mummy-sees-the bus*], [*Mummy-sees-Daddy*] or [*I-see-the house*]. In contrast, triadic representations place an embedded element in the third slot and *Self* in the first slot in order to specify that *Agent* and *Self* are attending to the same object. An example of this would be [*I-see-/Mummy-sees-Daddy/*].

Having laid out this description, Baron-Cohen then posits the existence of two specific modules in the brain: the Eye Direction Detector (EDD), which is capable of building dyadic representations through the use of precise geometric computations, and the Shared Attention Module (SAM), which builds triadic relations out of the dyadic representations it obtains from EDD, and assigns terms like “want” or “goal” using output from the Intentionality Detector (ID). Finally, a fourth module, the Theory of Mind Mechanism (ToMM), provides a mechanism for tying together our knowledge of mental states into a coherent whole. Baron-Cohen believes this mechanism uses experience to convert the representations of the SAM into mentalistic attributions, and is responsible for being able to say things like “John thinks Elvis is alive” (Baron-Cohen, 1995; Scassellati, 1999).

The Baron-Cohen model is not perfect. The attempt at a formal, symbolic description of the elements involved in sharing attention requires that a distinction between dyadic and triadic representation be made; however this distinction turns out to be made rather arbitrarily. For instance, separating [*She-sees-the bus*] from [*I-see-/she-sees-the bus*] implies first that in the dyadic relation it is possible to make that attribution at the earliest EDD level (but how can I know what she is seeing unless I put myself in her shoes?), and second that representing *Self* explicitly is necessary and even different from the dyadic representation (do I actually see something and then later *judge* that I am seeing it?) One of the problems with this is that this places enormous computational requirements on the EDD module, and then makes only very weak statements about how the later computations are made.

I do not doubt, in fact, that there is some kind of EDD. Baron-Cohen (1995) even justifies this module by describing neurological research implicating the Superior Temporal Sulcus (STS) as the location of the EDD. Cells in the STS of the macaque monkey are known to be selective for faces and specific face orientations, and Baron-Cohen refers to work even suggesting that certain cells in STS are sensitive to specific gaze directions.

The difficulty with this Fodorian description is that may not be able to take into proper account the time-locked nature of shared attention, an element which dynamical systems theorists would argue is the key ingredient of understanding cognitive processes (van Gelder & Port, 1995). Shared attention requires an active process of creating a representation of the others' gaze, *the active monitoring of another's gaze*, and *the attempt to manipulate the other's gaze*. Thus, while the ability to make judgements about intentionality based on static images of faces with a fixed gaze is a reasonable test of one of the results of a developed shared attention mechanism (a test described in Baron-Cohen, 1995), the real test lies in determining gaze and intentionality in a real-time setting.

Described this way, it seems possible that the SAM may not be a *thing in itself*, but rather an emergent property of the ability to incorporate many pieces of information into attention, and the prioritization of others' gaze as worthy of active monitoring. That is, it may simply indicate processing power and certain kinds of memory use, and not a specifically new skill. Using the result of the EDD is a then a different kind of task – it involves prioritizing this property of the world as important and useful in several different classes of activity, and it may require that one holds several pieces of information in memory: both the position of the referent while computing other's gaze, and the direction of other's gaze while computing the position of the referent. Actively holding these things in memory and updating them when necessary is a skill that requires both memory and an ability to carry out a single activity for an extended period of

time.

7 The Advantages of a Robotic Implementation

Although Baron-Cohen has made extensive use of experiments with autistics to test aspects of his model, it may never be possible to completely test his or any other model of shared visual attention directly in behaving humans. However, by implementing a model of visual attention in a robot, it may be possible to make tests of the model that otherwise would have to be left entirely to speculation.

In this spirit, researchers in the MIT Media Lab have been attempting to implement the Baron-Cohen model in an upper-torso humanoid robot called ‘Cog’. The attempt to create Cog has forced many of the questions of shared visual attention to be examined extremely closely. Of immediate importance is the question of understanding what kinds of perceptual features are most useful in determining the direction of the head and eyes (in implementing the EDD). Development of this robot allows the testing of many of the theories of what features are important that were described in section 5. Already many of these features are used in Cog’s visual system. Movement, color, and contingency are important aspects of his shared-attention skills, and implementation of these have allowed for the construction of several interesting behaviors, such as imitation of head-nodding by humans and other face-possessing objects.

Work is not complete on Cog, but the hope is that the same kinds of evaluation tools for joint attention mechanisms that are used on humans (such as the Vineland Adaptive Behavior Scales and the Autism Diagnostic Interview) can be used on him as well. If the behavior Cog produced was realistic for a normal human, then inhibiting specific modules would be a good test of the model to see if they produce behaviors similar to those of autistic children.

Although an implementation of the robot that produced realistic behavior

would not be able to make specific claims about the biological reality of the model, it would still hopefully lend insights into the computational requirements of making *any* implementation of shared visual attention work, and this insight could help to guide further investigations into the actual biological nature of these behaviors. Furthermore, it would be also be able to tell us that the model itself is internally consistent; if the behavior were wrong, it might point out to us where the model has inconsistencies that we might not have been able to observe if we were not able to vary the internal variables independently.

8 Conclusion

This account of shared visual attention has gone in the opposite direction from traditional cognitive science. Rather than going from the inside out, beginning with internal representations and then attempting to show how these representations allow one to interact with the world, this paper has gone from the outside in, beginning with social interaction and then working its way to successively lower levels right down to physical implementation. This direction for research is important, because it allows one to first have a very clear understanding of the actual behavior being produced and the interface between the outside world and the inside world before attempting to explain what lies inside the head. While not much of a biological account was not provided in this paper, there is plenty of evidence to make one, given adequate space and time.

This analysis is merely a first look at the growing body of literature on this subject, but it provides some suggestions for what the important areas of research in this subject are. It seems to me that this is one of the few areas in which something meaningful can be said at all levels of analysis and for which a single, coherent story can be told. This makes it a good beginning for creating a single story describing human behavior and a goal that other areas of cognitive science might strive for.

References

- Ardouino, P., & Gould, J. (1985). Is tonic immobility adaptive? *Animal Behavior*, *32*, 921-922.
- Baron-Cohen, S. (1995). The eye direction detector (edd) and the shared attention mechanism (sam): Two cases for evolutionary psychology. In C. Moore & P. J. Dunham (Eds.), *Joint attention: Its origins and role in development*. Hillsdale, NJ: Erlbaum.
- Butterworth, G. (1991). The ontogeny and phylogeny of joint visual attention. In A. Whiten (Ed.), *Natural theories of mind: Evolution, development, and simulation of everyday mindreading*. Oxford England: Blackwell.
- Butterworth, G. (1995). Origins of mind in perception and action. In C. Moore & P. J. Dunham (Eds.), *Joint attention: Its origins and role in development*. Hillsdale, NJ: Erlbaum.
- Cole, M. (1985). The zone of proximal development: where culture and cognition create each other. In J. V. Wertsch (Ed.), *Culture, communication and cognition: Vygotskian perspectives*. Cambridge: Cambridge University Press.
- Corkum, V., & Moore, C. (1995). Development of joint visual attention in infants. In C. Moore & P. J. Dunham (Eds.), *Joint attention: Its origins and role in development*. Hillsdale, NJ: Erlbaum.
- Corkum, V., & Moore, C. (1998). The origins of joint visual attention in infants. *Developmental Psychology*, *34*(1), 28-38.
- Deak, G. O., Flom, R. A., & Pick, A. D. (2000). Effects of gesture and target on 12- and 18-month-olds' joint visual attention to objects in front of or behind them. *Developmental Psychology*, *36*(4), 511-523.

- Grover, L. (1988). *Comprehension of the manual pointing gesture in human infants*. Unpublished doctoral dissertation, University of Southampton, England.
- Hutchins, E. L. (1995). *Cognition in the wild*. Cambridge, MA: MIT Press.
- Hutchins, E. L., & Strum, S. C. (1995). *Ontogeny of social skill* (Unpublished research proposal to the NSF No. 95-6006144-W). La Jolla, CA: University of California, San Diego, Department of Cognitive Science.
- Johnson, C. M. (1999). *Distributed primate cognition: a response to Tomasello and Call's Primate Cognition*. (Manuscript)
- Johnson, S., Slaughter, V., & Carey, S. (2000). Whose gaze will infants follow? the elicitation of gaze-following in 12-month-olds. *Developmental Science* (in press).
- Marr, D. (1982). *vision: A computational investigation into the human representation and processing of visual information*. Freeman.
- Moore, C., Angelopoulos, M., & Bennett, P. (1997). The role of movement in the development of joint visual attention. *Infant Behavior and Development*, 20(1), 83-92.
- Moore, C., & Dunham, P. J. (Eds.). (1995). *Joint attention: Its origins and role in development*. Hillsdale, NJ: Erlbaum.
- Movellan, J. R. (1986). Perception of directional attention. *Presented at the International Conference on Infant Studies*.
- Scaife, M., & Bruner, J. S. (1975). The capacity for joint visual attention in the infant. *Nature*, 253, 265-266.
- Scassellati, B. (1999). *Imitation and mechanisms of joint attention: A developmental structure for building social skills*.

Scheffer, R. (1984). *The child's entry into a social world*. New York: Academic Press.

van Gelder, T., & Port, R. F. (1995). It's about time: an overview of the dynamical approach to cognition. In R. F. Port & T. van Gelder (Eds.), *Mind as motion: explorations in the dynamics of cognition*. Cambridge, MA: MIT Press.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.