Mechanisms of intentional joint visual attention

Takeshi Konno

Abstract: People communicate with others via intention. This is likewise true for the primitive behavior of joint visual attention: directing one's attention to an object another person is looking at. However, the mechanism by which intention, a kind of internal state, causes that behavior is unclear. In this paper, we construct a simple computational model for examining these mechanisms, and investigate mechanisms for categorizing visual input and for recalling and comparing between these categories. In addition, we lay out some interaction experiments involving a human and a robot equipped with the constructed computational model, to serve as a platform for verifying the intentionality demonstrated by these mechanisms.

Keywords: Intention, Joint Visual Attention, Computational Model, Human-Robot Interaction.

1. *Introduction*

A person will naturally turn their attention to an object another person is looking at. This phenomenon is called joint visual attention (Butterworth, Jarrett 1991). Joint visual attention is the actualization of sharing an object under attention with another person; this ability is considered highly important for aiding social communication (Frith 1989) and vocabulary acquisition in humans (Tomasello 2003). Gaze behavior occurs almost entirely automatically, but many believe that even if it is reflexive at first, the behavior requires some sort of understanding of what the other person is paying attention to, since the attended object is shared. Tomasello (2000) writes that the understanding of the other as an intentional being like the self is a uniquely human cognitive competency. Consider two people facing each other across a dinner table with a salt shaker between them. When one looks at the salt shaker in front of him, intending to have the other hand it to him, the other may do so after understanding this intention. It is this – the other person's understanding of the first's object-oriented gaze as reflecting intention – that many

ISSN (print) 2281-8618-ETS

believe forms the foundation of uniquely human communication. Tomasello (2000: 72) continues by arguing that the understanding of others as an intentional agent is first prompted in infancy, when the infant's own actions become intentional.

The infant starts to experience joint visual attention with others in this stage of the developmental process; how they understand the intentions of others is contingent on how those intentions resemble their own. We return to our previous example, when the other man saw the first man focusing on the salt shaker. If that were me, the other man deduces, I could be looking at a salt shaker wanting someone to hand it to me. If we assume this to be true, we must clarify what it means for a person to become intentional. Anscombe (1957) writes that while an intention, such as the first man's intention to have the second man hand him a salt shaker in the example above, is indeed the reason for the behavior of looking, it is not the cause of this behavior *per se*. Philosophical debate continues to this day on the extremely difficult topic of defining an intention from observed behavior (Davidson 2001; Dretske 1997; Millikan 2004; Fodor 2008). Identifying an internal state as the cause of a behavior amounts to and constitutes the same endeavor as elucidating the internal mechanisms that produce that behavior. Progress in neuroscience is arguably indispensable to any investigation of the physical mechanisms behind intention, ranked as it is among the higher-order cognitive functions. Even if so, it shall be necessary to construct a model for advancing the discussion at computational, algorithmic, and representational levels (Marr 1982, Kaneko and Tsuda 1994, Hashimoto *et al.* 2008). Therefore, in this paper we attempt to construct a computational model for the phenomenon of joint visual attention: specifically, for mechanisms that generate the behavior of the self looking at an object intentionally. Through this work, we examine the configuration requirements of mechanisms that could produce intentional behavior. One could easily remark that joint visual attention is typically a reflexive action, and that the entity of 'intention' can be considered to be no more than an *a posteriori* interpretation of a behavior. In this paper however, we assume that having an internal state, which we might call a *particular goal*, is a prerequisite for and the cause of joint visual attention.

2. *Construction of the computational model*

Several constructive studies have dealt with joint visual attention, with many studies utilizing not only computational models but also robots. Consider a joint visual attention model using a robot. One basic task for the robot would be to detect the direction of human's gaze, and then shift its own gaze (field of view) to that direction. An important component of such a model is to direct the robot to an object not already within its field of view, i.e., for the human's line of sight to function as a signal pointing in that direction to some type of object. Many studies configure the relationship between vision and motion components *ab initio*, and incorporate the model into a robot (Breazeal and Scassellati 2000; Kozima 2002). However, what is generated cannot truly be called "purpose"; if the robot's sensory input is directly linked with the motor output that moves its field of view, nothing is generated there that we can call "purpose." One can conclude that what such systems generate is actually a reflexive action.

One review article by Kaplan and Hafner (2006) summarizes the efforts of constructive research to investigate the process by which the joint visual attention of infants develops into the joint visual attention involving sharing intentions with others. In the article, depicted alongside findings of cognitive developmental psychology is the process by which infants come to understand the intentionality of others, informed by its similarity with their own intentions. Kaplan and Hafner (2006: 139) note that the problematic "intention" in this process is an *action plan*, P, for reaching a *goal*, G, from the *initial state*, S; that plan includes a *particular goal* and a *means*. What are the components of a particular goal and means? Unfortunately, no explanations or models have been offered for such a system to date. We will now discuss this mechanism.

Few constructive studies are concerned with how to make an agent learn relationships between the visual and motor components of joint visual attention (Triesch *et al.* 2006; Nagai *et al.* 2003; Matsuda and Omori 2001). Certainly, the learning process includes an internal evaluation of the results of the action, alongside trial-and-error movements for reaching the goal. Therefore, one could treat that goal as the purpose. Piaget (1952) observed that infants begin to show object-oriented behaviors at 8 months of age, by removing obstacles to objects. This would appear to constitute an intentional action. However, previous computational models choose among different action options (e.g., shifting the field of view) corresponding to different visual input (e.g., the human's gaze) by trial-and-error, and evaluate whether the actions have resulted in the visual input reaching the goal state (i.e., directing the line of sight at the designated object). We can summarize the input and output in the schematic diagram in Fig. 1. There is no way for the goal state to trigger any action options directly.

Triesch *et al.* (2006) adopt this learning framework in their own computational model, in which learning results in infant agents referencing the gaze

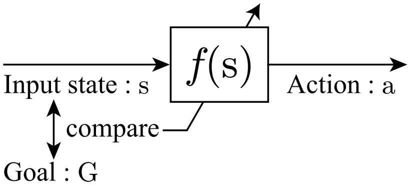

Fig. 1. A mechanism for behavior learning by trial-and-error.

direction of another person, as if with the "purpose" of learning where a toy is. This is because the model has the infant's experience finding a toy of interest in the other person's line of sight, and the action selection algorithm incorporates the fact that the other person's gaze is an effective source of directional information pointing to the toy. However, this action in the model actually consists of a chain of state transitions due to actions that are selected in direct response to input states; selecting this action is not necessarily grounds for concluding the presence of purpose.[1] In other words, even if an agent's actions generated by a learning process appear to reflect a specific goal, the action-selection system is not necessarily influenced by any given purpose. Moreover, there is no internal state that acts as a direct cause of the action selection, complicating the notion that we utilize an observed action's similarity to our own actions as the basis for deducing the purposes of others. This is because the system can only directly infer sensory input states from actions, not the goal state.[2] What could be the components of a system featuring an internal purpose that caused the action selection?

Fig. 2 shows the simplest system we could think of. This system generates an internal particular goal, G', in response to sensory input, s, and selects an action, a, according to that particular goal state. In this system, joint visual attention produces a particular goal state based on the gaze of others. The agent's field of view is moved according to that particular goal state, eventually coming to land on an object located along the other person's line of sight. For a particular goal, G', we must consider how it will be generated along with an action selection function, f (G'). This action selection function does not accept

[1] Nagai *et al.* (2003)'s research differs from Triesch *et al.* (2006)'s model, by using artificial neural networks to connect sensory input with behavioral output. Learning from episodes of joint visual attention for objects within the visual field, their robot comes to learn how to jointly attend to objects located outside it. However, its essential mechanisms are similar to the computational models of Matsuda and Omori (2001) and Triesch *et al.* (2006), leaving the outstanding issue of purpose not being the direct cause of actions.

[2] Of course, a model could have interactions between actions and the goal state for evaluating transition states. Considering these relationships separately enables the model to deduce the other person's goal state.

gaze direction of the human as input; therefore, its nature must accordingly be different from joint visual attention *per se*. In this paper, we thus consider the action of merely focusing one's gaze on an object that is reflected in the field of view. Infants are observed to acquire this behavior, called *visual orientation*, by around 3 months of age (Atkinson *et al.* 1992).

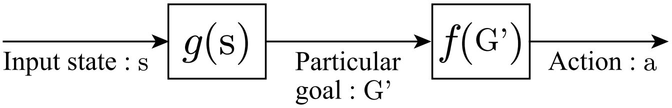

Fig. 2. Generating mechanism of the particular goal.

We create a simple model in order to examine a computational-theoretical implementation of visual orientation (Fig. 3). A horizontal row of discrete squares represents the world as seen by the infant agent. The agent's field of view measures three squares, (p_L, p_C, p_R), with its focal center in the middle square. An object is also placed in the row: this could be the face of the other person, characterized by a gaze direction of left (←) or right (→), or a toy, characterized by a circular (●) or rectangular (■) shape. When one of these object enters the agent's field of view, both the specific square within the field of view in which the object appears, and feature-related data are input into the system, s_V. For example, the input, s_V, for a left-facing human face appearing in the left-side field-of-view square would be (←, p_L), while a circular toy appearing in the center field-of-view square would be (●, p_C). To simplify the model, it does not account for cases of multiple objects in the agent's field of view.

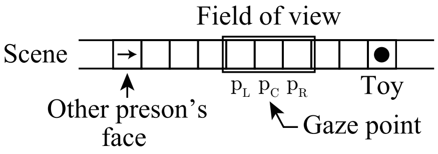

Fig. 3. The environment of the computational model.

Visual orientation is modeled as the act of the agent fixing their gaze on another person's face or toy reflected in the field of view, i.e., regarding an object in the center of the field of view. We allow for the agent to select the actions, a, of shifting their line of sight one square to the left or right. Achieving visual orientation depends on the agent's selection of appropriate behavior in response to the input, s_V, and position information, p. One could conceivably create a model in which the agent's selection of actions is informed by reinforcement learning (Sutton, Barto 1998). However, we decided to have the model assume that the agent is already capable of the behavior. If we model visual orientation by the action selection function, *f* (G'), in Fig. 2, then we must ask how the model should generate its particular goal state, G', in response to the other's gaze. We can think about this using Fig. 4, a step-by-step illustration of joint visual attention behavior until the agent focuses his or her line of sight on a toy originally placed outside its field of view.

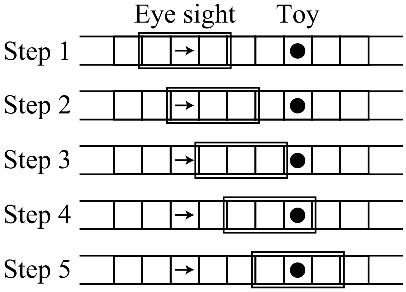

Fig. 4. Action sequence of joint visual attention.

The behavior begins from a state in which a right-facing face is reflected in the center of the agent's field of view (Step 1). The agent first shifts his or her line of sight to the right (Step 2). If the agent chooses visual orientation as the action at this point, his or her line of sight would return to the left in the next step. Nothing is reflected in the agent's field of view when it shifts it to the right again (Step 3). In this state as well, the agent must continue to shift its line of sight to the right. The toy finally enters the agent's field of view in the right-side square after shifting further still to the right (Step 4). If the agent chooses visual orientation at this point, the agent's gaze will shift to focus on the toy (Step 5).

There are a few requirements for implementing the behavior described above. The agent must respond to the input of gaze direction, $s_V = (\rightarrow, p_C)$, in Step 1 by shifting his or her field of view to the right. At this time, the agent's field of view moves one square to the right if the particular goal, G', of (●,[3] p_R) (i.e., the toy being located in the right square) is output by the function for generating the particular goal, *g* (s). In other words, the other party's gaze causes the agent to *recall* the toy. If the agent continues to recall the toy in Steps 2 and 3 as well, their line of sight will likewise continue to shift to the right. In Steps 4 and 5, however, we must set the particular goal, G', of the toy being present

[3] The toy needs not be circular in shape.

in the agent's field of view. If not, their field of view would pass over the object. What Steps 2 to 5 involve is a categorical comparison between the particular goal stored in memory, G', and the visual input, s_V. This category distinguishes the kind of object, without respect to its location. In Steps 2 and 3, the toy in memory is compared with an empty field of view (i.e., a state in which nothing is reflected); in Steps 4 and 5, the toy in memory is compared with the toy seen in the field of view. The particular goal generation function, *g* (s), continues to compare the object in memory with the visual input until they belong to the same category.

We can summarize our above system for joint visual attention via the application of visual orientation as follows. The system must first create and internally retain an association between two objects of different categories (i.e., the gaze of another actor and a toy), and internally retain a recalled state (i.e., the toy in memory) until something belonging to the same category enters its visual input. One important feature of this model is that the agent needs to "recall" a toy in the other actor's line of sight. We believe the adoption of recall as a necessary function is suitable for the model; humans can recall an object if they accumulate episodes of experiencing it visually within the field of view using visual orientation. Infants should become able to recall a toy being the focus of another actor's gaze if this relationship is selectively reinforced by experiences of interesting toys located in others' lines of sight and within the infant's own field of view. Another important feature of the model is that the recalled object need not continue being the same specific entity. Generalization across many experiences means that the recalled object can enter a "wild-card" state; when performing the behavior of joint visual attention in this case, the agent will still act to see what the object is if the only specific input remaining is the direction to its location. Conversely, when the agent is recalling a specific object, they will focus their attention on it even if several other objects are present in the gaze direction, and moreover, will continue to search if it is not found in the gaze direction. The agent's stance towards the object is thus variable, being dependent on the specificity of the recalled object. The hypothetical, recalled object is maintained internally in every possible state in this model. It serves as the agent's particular goal, or *what they are trying to see*; visual orientation serves the means to achieve this goal.

3. *Implementing the computational model in a robotic platform*

The computational model constructed in the previous section is not new; we have employed it as a computational model for specific behaviors (Konno and Hashimoto 2006), and have implemented it in a robotic platform to run human-robot interaction experiments (Konno and Hashimoto 2010). In the experiment, we had a human and a robot face each other across a table (Fig. 5). The robot we constructed had simple functionality and appearance, combining a stereo camera corresponding to the eyes, with a stand that could pan and tilt the camera. We thought it would be better to keep its appearance very simple, as the main focus of the experiment was to demonstrate a functional application of the mechanisms captured by the computational model. Sheets of paper with the numerals 1 to 12 written on them were placed on the table. In the experiment, the person looked at a number in front of them, and the robot jointly focused on that number based on the person's gaze. The close positioning of the numbers next to each other made it difficult for the robot to determine which number the human was looking at based on their gaze. Human-robot interactions were driven under these conditions by two different mechanisms. The first mechanism was to generate the behavior of directly following a human's gaze. After shifting its field of view a certain distance to follow the human's gaze, the robot focused on the number closest to the field's focal center. This means that although the robot's movements are certainly swift by and large, the last focusing step represents the fine-tuning behavior of calibrating the field of view. The robot was configured to return to the human's face after focusing on the number for a certain length of time.

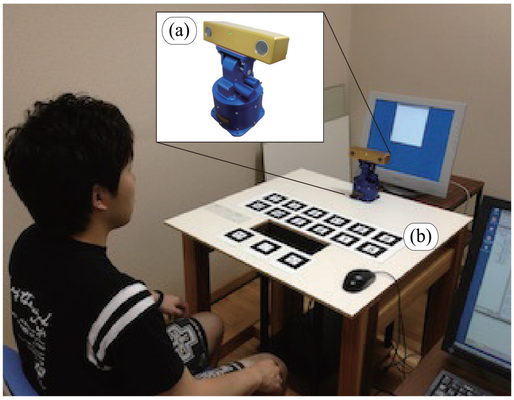

Fig. 5. Experimental Environment. A Robot (a) and twelve numbers (b).

The second mechanism was to recall a number based on the human's gaze, and to move the field of view to that number's spatial location. This robot immediately moved its field of view toward the recalled number. The numbers recalled by the robot were determined in advance, based on a frequency distribution of the numbers cumulatively encountered by the robot in tests of the first mechanism with several different human participants. This means a given number "recalled" by the robot is determined probabilistically, chosen from among the numbers present in the human's gaze direction. The robot accurately focused on any given number it recalled, because each number is associated with accurate position information in the mechanism. Participants in this experiment were told in advance by the experimenters that the two robots they interacted with ran according to different mechanisms.

Participants thus interacted with robots whose movements differed yet had the same appearance. Unfortunately, however, our human participants did not comment on significant differences between these two robots having different mechanisms. Certainly, many participants commented in the individual interviews and questionnaires that they had felt the presence of intention in the robots' behavior. We can, however, divide the perceived intentions according to the mechanism: for the second robot, participants reported feeling intentionality in the action of rapidly turning its gaze to the object, but for the first robot, intentionality was felt for the action of fine-tuning the field of view to focus on the number in the person's line of sight. The experiment was additionally complicated from the outset because of the tendency for humans to perceive intention in response to any behavior.

The results force us to admit that it is extremely challenging to quantitatively clarify differences in intentionality through behavioral observation. It seems almost impossible to verify the presence of intentionality in such a prototypical behavior. However, the human-robot interaction experiments give us an idea of whether the behavior exhibited by the constructed computational model authentically reflects intentionality. We consider it essential to investigate mechanisms by which the higher-order faculty of intentionality is shared, with such human-robot interaction experiments as a verification platform.

4. *Discussion: development of a computational model for joint visual attention based on shared intentionality*

In this paper, we investigated the mechanisms by which an internal state could cause the behavior of joint visual attention. These mechanisms were implemented in a system in which the agent recalls a previously seen object in response to the gaze of another person. This system's plan is composed of a *particular goal* – the recalled object, i.e., the object to be viewed – and a *means*, the pre-existing ability of visual orientation. Visual orientation uses only the object's position information in this framework. In contrast, the function producing the particular goal must make associations and comparisons between categories that distinguish the type of the object. Different functionalities emerge because each individual function deals with different classes of data.

Categorization and generalization are important in this regard. The input state prepared for the computational model constructed in this paper consists of two components: gaze direction, (←, →) or toy shape, (●, ■), and position location, (p_L, p_C, p_R). The agent does not require the first component for visual orientation; however, the input state must distinguish between the categories of human face and toy in order to realistically model joint visual attention. Human vision can capture diverse kinds of information, but realistically modeling a function that differs from existing functions seems to require separating this information into different categories. In addition, our computational model supposes that the agent recall a toy of a specific shape, but it is important that the agent be able to make two kinds of generalizations based on the object: categoric generalization and wild-card generalization. In this context, categoric generalization could provide that the recalled object can be anything as long as it is a toy. Wild-card generalization could provide that the recalled object need not be a specific thing, merely *something*. In this situation, the agent does not know what that something is, but would still shift his or her field of view according to the other's gaze direction. It may indeed be more natural to think about communication as beginning from this generalized state, and becoming more specific through the two parties exchanging actions.

The problem is, people sometimes shift their view immediately to follow the gaze of others. The mechanism of action here likely differs from the mechanism of action of following the gaze of another based on some kind of recalled state. To account for this difference, we endorse the theory that multiple mechanisms related to behavioral decisions are present in humans. The dual-process theory proposed by Keith (2004) holds that processing mechanisms in humans fall broadly into two categories: automatic, speedy parallel processing, and analytic, slow series processing. We believe that the two mechanisms presented in this paper are the most primitive versions of these two processing mechanisms. If joint visual attention is limited to a human's capability to follow a gaze, there would be no use for any recall mechanism. In this regard, it is important that we investigate whether a series-type processing mechanism for recall can be developed sufficiently to model intention sharing as described by Tomasello and Carpenter (2007).

What kind of process would need to be developed first? We hypothesize that the agent model would first need to be able to infer the other person's *particular goal* based on similarities of relationships between his or her own *particular goal* and *means*.[4] In the series-type recall mechanism, we propose that actions are directly caused by the agent's recalled internal state. The proposed mechanism means that internal state can be understood as the cause of an action, if the actor had inferred the other person's internal state from the action they observed. For example, when the agent observes that the other person is looking at a certain object, the internal state that appears to be promoting the viewing behavior in the other person enters the agent's mind. This internal state is not merely the other person's desire, nor is it the final goal. We believe this is the primitive version of the action plan, P, identified by Kaplan and Hafner (2006), and consider intention to constitute the set of a particular goal and a means contained in this action plan. If the agent forms his or her own particular goal based on the inferred particular goal, the agent would determine whether he or she is seeing the object based on the object the other intends to see. In this event, could we reasonably say that the agent's behavior is joint visual attention based on his or her understanding of another human's purpose?

Finally, let us consider joint visual attention based on the shared intentionality (Tomasello, Carpenter 2007). When the agent has turned its attention to the same object the other person was looking at, comprehends it as such, and are looking together, can we say that the experience of seeing the object is being shared by the two people? This statement is probably insufficient, because even if the first person knows that the second person is looking at that object, the second person's understanding is missing the information that the first person knows this. The scenario requires a nested structure of intention between the self and the other. This nested structure is widely mentioned in philosophical discussions of the mind (Dennett 1987, Colombetti 1993, Sperber and Wilson 1995, Carston 2002), and continues to be passionately pursued in theories of mind (Premack and Woodruff 1978, Premack 1988) in cognitive developmental psychology (Wimmer and Perner 1983, Emery 2000, Saxe and Young 2013). Attempts at constructing computational models based on these pursuits have naturally been made in the field of artificial intelligence as well. However, no artificial entities have yet been created in which humans can perceive shared intention, even subjectively. Let us for now suppose that we are not mistaken

[4] The supposed mechanism presumes that an agent cannot directly know the particular goals of another person. However, one basic phenomenological theory holds that the particular goals of others are perceived directly through one's own body (Husserl 1960: 33). In this theory, the assumed mechanism is grounded in the fact that the agent and the other share the same physical form. Mirror systems (Rizzolatti and Craighero 2004) observed in neural networks in the brain have drawn attention in recent years as proof of the existence of such mechanisms. It is essential to consider a wide range of possibilities with regard to mechanisms for understanding the intentionality of others.

to assume the existence of a nested structure relating the self and others that incorporates the pair of the particular goal and means. What is this cause of? The computational model presented in this paper suggests the importance of mechanisms for categorization and generalization when the agent forms his or her particular goal. For a model to exhibit a different function from the means (i.e., visual orientation), it needs to create categories different from the state acted on by the means, and to make associations between those categories. In addition, the model's ability to generalize a specific object allowed us to discuss how the agent's intentional stance varies and changes with respect to objects. Accordingly, in order for the computational model presented in this paper to evolve and represent a true state of shared intention, it needs to be configured so that the particular goal forms a nested structure between the self and others, while continuing to incorporate mechanisms for categorization and generalization of the particular goals of the self and others.

Takeshi Konno
Kanazawa Institute of Technology,
Information and Communication Engineering, Dept. of Electronics,
konno-tks@neptune.kanazawa-it.ac.jp

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP15H03151, JP23320002, and JP24700258.

References

Anscombe, G.E. Margaret, 1957, *Intention*, Basil Blackwell, London.

Atkinson, Janette, Bruce Hood, John Wattam-Bell, Oliver Braddick, 1992, "Changes in infants' ability to switch visual attention in the first three months of life", in *Perception*, 21: 643-653.

Breazeal, Cynthia, Brian Scassellati, 2000, "Infant-like social interactions between a robot and a human caretaker", in *Adaptive Behavior*, 8: 49-74.

Butterworth, George, Nicholas Jarrett, 1991, "What minds have in common is space: Spatial mechanisms serving joint visual attention in infancy", in *British Journal of Developmental Psychology*, 9: 55-72.

Carston, Robyn, 2002, *Thoughts and Utterances: The Pragmatics of Explicit Communication*, Blackwell, London.

Colombetti, Marco, 1993, "Formal semantics for mutual belief", in *Artificial intelligence*, 62, 2: 341-353.

Davidson, Donald, 2001, *Essays on Actions and Events*, 2nd ed., Oxford University Press, Oxford.

Dennett, Daniel C., 1987, *The Intentional Stance*, MIT Press, Cambridge MA.

Doherty, Martin, 2008, *Theory of Mind: How Children Understand Others' Thoughts and Feelings*, Psychology Press.

Dretske, Fred I., 1997, *Naturalizing the Mind*, MIT Press, Cambridge MA.

Emery, Nathan J., 2000, "The eyes have it: the neuroethology, function and evolution of social gaze", in *Neuroscience & Biobehavioral Reviews*, 24, 6: 581-604.

Fodor, Jerry A., 2008, *LOT 2: The language of thought revisited*, Oxford University Press, Oxford.

Frith, Uta, 1989, *Autism: Explaining the Enigma*, Blackwell, London.

Hashimoto, Takashi, Takashi Sato, Masaya Nakatsuka, Masanori Fujimoto, 2008, "Evolutionary constructive approach for studying dynamic complex systems", in Petrone, Giuseppe, Giuliano Cammarata eds., *Modelling and Simulation*, I-Tech Education and Publishing: 111-136.

Husserl, Edmund, 1960, *Cartesian Meditations*, Dorion Cairns trans., Springer Netherlands, The Hague.

Kaneko, Kunihiko, Ichiro Tsuda, 1994, "Constructive complexity and artificial reality: an introduction", in *Physica D*, 75: 1-10.

Kaplan, Frederic, Verena V. Hafner, 2006, "The challenges of joint attention", in *Interaction Studies*, 7, 2: 135-169.

Keith, E. Stanovitch, 2004, *The Robot's Rebellion: Finding Meaning in the Age of Darwin*, University of Chicago Press, Chicago.

Kozima, Hideki, 2002, "Infanoid: A babybot that explores the social environment", in Dautenhahn, Kerstin, Alan H. Bond, Lola Canamero, Bruce Edmonds eds., *Socially Intelligent Agents: Creating Relationships with Computers and Robots*, Springer, Boston: 157-164.

Konno, Takeshi, Takashi Hashimoto, 2006, "Developmental construction of intentional agency in communicative eye gaze", in *Proceedings of the International Conference on Development and Learning*, ICDL06, Indiana Univ., USA, May 31st-June 3rd: 6 pages.

Konno, Takeshi, Takashi Hashimoto, 2010, "An experiment with human-robot interaction to study intentional agency in joint visual attention", in *Proceedings of the 9th IEEE International Conference on Development and Learning*, ICDL9, Univ. of Michigan, USA, Aug. 18th-21st: 2 pages.

Marr, David, 1982, *Vision: A computational investigation into the human representation and processing of visual information*, MIT Press, Cambridge MA.

Matsuda, Goh, Takashi Omori, 2001, "Learning of joint visual attention by reinforcement learning", in *Proceedings of the International Conference on Cognitive Modeling*, ICCM2001, George Mason Univ., USA, July 26th-28th: 157-162.

Millikan, Ruth Garrett, 2004, *Varieties of meaning: the 2002 Jean Nicod lectures*, MIT Press, Cambridge MA.

Nagai, Yukie, Koh Hosoda, Akio Morita, Minoru Asada, 2003, “A constructive model for the development of joint attention”, in *Connection Science*, 15, 4: 211-229.

Piaget, Jean, 1952, *The Origins of Intelligence in Children*, Margaret Cook trans., Norton, New York.

Premack, David, Guy Woodruff, 1978, “Does the chimpanzee have a theory of mind”, in *Behavioral and Brain Sciences*, 1, 4: 515-526.

Premack, David, 1988, “‘does the chimpanzee have a theory of mind?’ revisited”, in Byrne, Richard W., Andrew Whiten eds., *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes and Humans*, Oxford University Press, New York: 160-179.

Rizzolatti, Giacomo, Laila Craighero, 2004, “The mirror-neuron system”, in *Annual review of neuroscience*, 27: 169-192.

Saxe, Rebecca, Liane Young, 2013, “Theory of mind: How brains think about thoughts”, in Ochsner, Kevin N., Stephen Kosslyn eds., *The Oxford Handbook of Cognitive Neuroscience: Volume 2: The Cutting Edges*, Oxford University Press, New York: 204-213.

Sutton, Richard S., Andrew G. Barto, 1998, *Reinforcement Learning*, MIT Press.

Sperber, Dan, Deirdre Wilson, 1995, *Relevance: Communication and Cognition*, Wiley-Blackwell, London.

Tomasello, Michael, 2000, *The Cultural Origins of Human Cognition*, Harvard University Press, Cambridge MA.

Tomasello, Michael, 2003, *Constructing a Language: A Usage-based Theory of Language Acquisition*, Harvard University Press, Cambridge MA.

Tomasello, Michael, Malinda Carpenter, 2007, “Shared intentionality”, in *Developmental Science*, 10, 1: 121-125.

Triesch, Jochen, Christof Teuscher, Gedeon O. Deak, Eric Carlson, 2006, “Gaze following: Why (not) learn it”, in *Developmental Science*, 9, 2: 125-147.

Wimmer, Heinz, Josef Perner, 1983, “Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children’s understanding of deception”, in *Cognition*, 13, 1: 103-128.