A Computational Developmental Model of the Implicit False Belief Task

Vincent G. Berthiaume (Vincent.Berthiaume@McGill.ca)
Department of Psychology, 1205 Dr. Penfield Avenue
Montreal, Qc, H3A 1B1 Canada

Kristine H. Onishi (Kris.Onishi@McGill.ca)
Department of Psychology, 1205 Dr. Penfield Avenue
Montreal, Qc, H3A 1B1 Canada

Thomas R. Shultz (Thomas.Shultz@McGill.ca)
Department of Psychology and School of Computer Science, 1205 Dr. Penfield Avenue
Montreal, Qc, H3A 1B1 Canada

Abstract

Do children understand that others have mental representations, for instance, mental representations of an object's location? This understanding, known as a representational Theory of Mind (ToM) has typically been studied using false-belief (FB) tasks. Standard, verbal FB tasks test whether a child can use protagonists' beliefs to say that they will search for objects where they last saw them. Whereas children under 3.5 years typically fail the task and expect protagonists to search where objects are (expectation consistent with an omniscient ToM), older children expect protagonists to search where they last saw the objects (expectation consistent with a representational ToM). Recently, 15-month-olds were shown to succeed at a visual, implicit version of the task. We present a sibling-descendant cascade-correlation connectionist model that learns to succeed at an implicit FB task. When trained on twice as many true- as false-belief trials, our model reproduced the omniscient-to-representational transition observed in explicit tasks. That is, networks first had expectations consistent with an omniscient ToM, and after further training had expectations consistent with a representational ToM. Thus, our model predicts that infants may also go through a transition on the implicit task, and suggests that this transition may be due in part to people holding more true than false beliefs.

Keywords: False beliefs; theory of mind; connectionism.

Background

Do children understand that others have mental representations, for instance, mental representations of an object's location? This understanding – known as a representational Theory of Mind (ToM, Wimmer & Perner, 1983) – has been found, using explicit false belief tasks, to go through a developmental transition between 3 and 4 years of age.

In this task, children see a puppet named Sally put a marble in a basket that is next to a box. While Sally is gone, another puppet, Anne, moves the marble from the basket to the box, thereby leaving Sally with the false belief that the marble is still in the basket. To predict that Sally will search for the marble in the basket, children must understand that she has a mental representation of the scene that is not consistent with reality (Dennett, 1978). A child under three years and eight months (Wellman, Cross, & Watson, 2001) will typically say that Sally will search in the box, an expectation that is consistent with an omniscient ToM – i.e., Sally will search in the actual location of the object. An older child will instead typically say that Sally will search in the basket, an expectation that is consistent with a representational ToM – i.e., Sally will search for the object in accord with her mental representation of its location.

Recently, 15-month-olds were shown to solve a visual, implicit version of the task (Onishi & Baillargeon, 2005). Because it avoids the complexities of language, the implicit task is more amenable to computational modeling. This paper introduces a connectionist model that learns to solve the implicit false-belief task and that reproduces the omniscient-to-representational ToM transition when training contained more true- than false-belief trials.

The Implicit False-Belief Task

Onishi and Baillargeon (2005) used a violation-of-expectation paradigm to show that 15-month-olds could pass an implicit, language-free version of the false-belief task. This paradigm uses looking time as a dependent measure of surprise. Infants – just like adults – look longer at the unexpected.

Infants were seated on their parent's lap and watched an actor performing actions with two boxes (one green and one yellow) and an object (Onishi & Baillargeon, 2005). First in familiarization trials, infants were shown the actor puting the object in the green box, as if to hide it there. Then on two trials the actor searched in the green box as if to retrieve the object (without actually revealing it) to convey to infants that she wanted it. Next, infants saw one of four belief-induction trials designed to cause the actor to hold either a true belief (TB) or a false belief (FB) that the object was either in the green or yellow box. For instance, in the TB-yellow trial, infants saw the actor watching as the object moved from the green box to the yellow box, thus causing the actor to hold a TB that the object was in the yellow box. By contrast, in the FB-green trial, another group of infants saw, as the actor was absent from the scene, the object move from the green into the yellow box, thus causing the actor to

have a FB that the object was still in the green box. The TB-green and FB-yellow trials were constructed with similar manipulations, the actor watching or not as the object was moved from (or stayed in) the green box. Finally, each infant saw one of two test trials in which the actor searched in either the green or yellow box. As can be seen in Figure 1, infants expected the actor to search according to her belief, whether true or false, and looked reliably longer when she did not do so. When the actor's belief (true or false) suggested that the object was in the green box, infants looked reliably longer when the actor searched in the yellow box (white bar is taller than gray bar in TB-green and in FB-green), while when the actor's belief suggested the object was in the yellow box, infants looked reliably longer when the actor searched in the green than the yellow box (gray bar is taller than white bar in TB-yellow and FB-yellow).

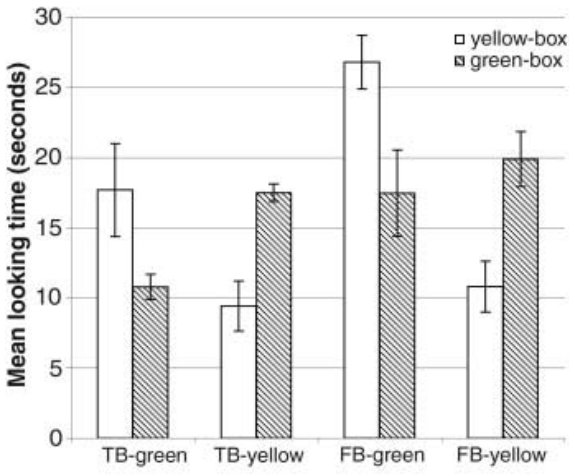

Figure 1: Infants' mean looking time and SE bars for the implicit FB task in the eight task conditions (four belief-induction by two test trials). From Onishi and Baillargeon (2005). Reprinted with permission from AAAS.

In sum, this work showed that 15-month-olds can succeed at an implicit false-belief task involving an approach goal. In contrast, not before 44 months are children able to reliably succeed at the verbal version of an approach task. Because it avoids language, the implicit task is more amenable to computational modeling and it is the task we chose to model. Before introducing our model, we review previous computational models of explicit FB tasks; no model of the implicit task has been proposed before. Although the explicit and implicit tasks differ substantially, comparisons between our model and previous models of the explicit task are possible because those models did not explicitly incorporate language.

Previous Models of False Belief Tasks

O'Laughlin and Thagard (2000)

O'Laughlin and Thagard's (2000) model of the standard FB task was based on the hypothesis that the inability to find coherence between concepts is what leads to poor performance. In their model, the authors connected elements roughly corresponding to propositions from the explicit FB task, e.g., "Sally puts marble in basket", etc., with either positive links between coherent elements or negative ones between incoherent elements. Depending on which values the experimenters assigned to the connections, the model made different predictions about where Sally would search. This model covered the pattern of data – predictions consistent with an omniscient ToM with some parameter values and predictions consistent with a representational ToM with others. However, because the patterns of connectivity for the two sets of expectations were built in by the programmers, it is not really a developmental model. The model did not go through the omniscient to representational ToM transition on its own.

Triona, Masnick and Morris (2002)

Triona, Masnick, and Morris' (2002) model used the ACT-R production system (e.g., Anderson et al., 2004) to model failure and success on a FB task (Perner, Leekam, & Wimmer, 1987). Triona et al. (2002) reproduced the observed transition in children's responses "by manipulating … the probability that the production [output] would achieve the goal" (p. 1045). When the probability parameter was low, the output was wrong (e.g., predicting Sally would search in the box), but when the parameter was high, the output was correct (e.g., predicting Sally would search in the basket). Once again, this model covered the experimental data through parameter manipulation by the experimenters, and is thus not a model that undergoes an autonomous developmental transition.

Goodman and colleagues (2006)

Goodman et al. (2006) built two causal Bayesian networks for the FB task. In the omniscient network, Sally's belief depended only on the marble's location, whereas in the representational ToM network, Sally's belief depended both on the marble's location and Sally's visual access to the marble's displacement. This difference made the omniscient network fail the FB task and the representational one succeed, but the extra connection also made the representational network more complex. Goodman et al.'s model suggested that it would be parsimonious for children to first use the simpler omniscient theory, but after accumulating evidence for its inadequacy, it would make sense to switch to the more complex representational theory. The model learned the posterior probability distributions of the search location from prior probability distributions, showed that expectations consistent with a representational ToM required more computational resources than do expectations consistent with an omniscient ToM, and switched from predictions consistent with an omniscient ToM to those consistent with a representational ToM. However, the architecture of both Bayesian networks were built by the experimenters, and by choosing which networks to include in the model, the set of possible transitions was

restricted to only 2; simple to complex (as the model selected) or complex to simple, thus showing an autonomous transition but one that was highly constrained.

In sum, previous computational models typically required substantial experimenter manipulation and involved restricted or no autonomous development. These issues are naturally addressed by our model, because it uses constructive neural networks.

Experiment

We conducted an experiment using fifty-six neural networks implementing the sibling-descendant cascade-correlation (SDCC, Baluja & Fahlman, 1994) constructivist algorithm, which has been successful in modeling numerous psychological phenomena (e.g., Shultz & Bale, 2006). Our model learns to succeed at the implicit false-belief task and transitions from omniscient to representational ToM expectations with additional training. It does not require parameter manipulation to make that transition but does require more true- than false-belief information in training. In effect, most of the time, beliefs *are* true (Leslie, German, & Polizzi, 2005), and this could potentially lead three-year-olds to expect others to hold true beliefs by default. Also, it may require more sustained attention to process or perceive a false-belief situation simply because there are more pieces of information to track for a longer time (previous location of the object, actor not looking as the object moves, etc.) compared to true-belief situations, in which infants could simply rely on their own knowledge of the object's location. For these reasons, networks were trained on twice as many true- as false-belief trials. When equal numbers of true and false belief trials were included during training, our model succeeded at the task without first going through a period of omniscient expectations.

Method

Initial Network Structure In everyday life, there are almost always multiple locations that an object can be. We trained networks using four locations. There is nothing special about four locations, but we wanted more than two for increased realism. Figure 2 represents the model in its initial state, i.e., with input and output units but without any hidden units. Constructivist networks such as SDCC are initialised without any hidden units, but they recruit them as more computational power is required in training.

Input The model's inputs represent the critical factors that govern searching behaviour and predictions made by an observer about that behaviour. The inputs encoded: (1) the first location of the object, (2) the second location of the object, (3) whether the actor was watching when the object was moved and (4) experimental context.

Four input units represented in which of the four locations the object started. These locations could be thought of as red, yellow, blue, and green boxes. Another four inputs represented where the object ended up. The object's location for a given time step was encoded by activating the location of the object but not the other three locations. One input unit encoded whether the actor was watching or not as the object moved. The tenth and last input encoded experimental context, a random value (between 0 and 1, selected from a uniform distribution) used to facilitate the network's stochastic training. As explained below, networks were trained on observations of behaviour that were not always correct in order to have training that is more like everyday observations. This type of stochastic training is problematic for deterministic neural networks because they cannot match different outputs to the same input. By adding an input node encoding a value randomized for each trial, much like different contexts for human experiences (time of day, etc.), our networks were able to learn the task.

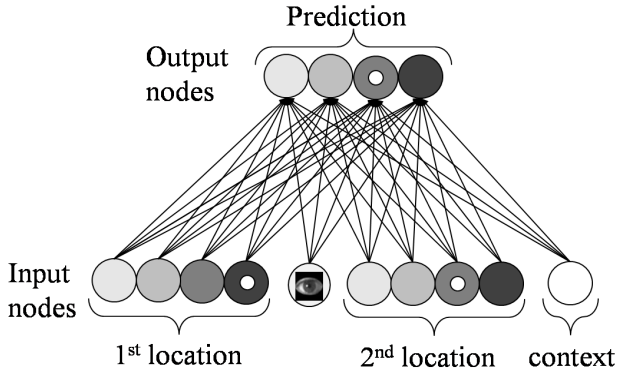

Figure 2: Schematic of an initial network used to model the learning of false beliefs. In the displayed event, the actor is looking (eye is present in the middle bottom unit) as the object moves from the fourth location to the third location.

Output Our model had 4 output nodes that represented in which, out of the 4 locations, the model predicted searching (see Figure 2). We simulated looking times by comparing the outcome the model predicted to the outcome that would be expected with a representational ToM, with larger discrepancy or error corresponding to longer looking times (Shultz & Bale, 2001).

The output nodes encoded the probability (between 0 and 1) that the actor would search in each location (asigmoid activation functions compressed the net input to an output unit into the range 0-1). For instance, if a network's output was red = 0.10, yellow = 0.70, blue = 0.05, green = 0.15, this was interpreted as the network expecting search in red 10% of the time, in yellow 70% of the time, in blue 5% of the time, and in green 15% of the time.

Training In everyday life, people probably do not always search for objects where they last saw them, because they forget, were distracted, etc. Therefore, networks were trained with observations of searching that were correct 18 times out of 21, or 85.7% of the time, to increase training realism. For example, if the actor should believe the object was in blue, the network was trained 85.7% of the time on correct trials (search in blue) and 14.3% of the time on

incorrect trials (search in red, yellow or green, equally at random).

Before training, SDCC networks do not have any hidden units, but only have direct connections between input and output units. During training, the network updates its weights, reducing error at the output. When error fails to decrease sufficiently, the network selects and recruits the one unit from a small number of potential hidden units that most reduces output error.

Training usually continues until the output error, i.e., the absolute difference between output activation and target output value goes below the score threshold (ST) parameter value (here kept at the default value of 0.4) for every output unit on every training pattern. However, since our networks were trained on stochastic observations, the output error never went below the ST (unless ST was set to 0.5, in which case networks only focused on one of the output values and did not learn the full training probability distribution). Therefore, training was not terminated using the ST, but instead it was stopped after networks learned the probability distributions of search in training (18 correct searches out of 21 for any given belief-induction trial).

Learning was assessed right before networks recruited each hidden unit, using chi-square ratios to test whether, for each condition, expected and observed frequencies were significantly different. Expected frequencies were the frequencies used in training, i.e., 18/21 correct searches. Observed frequencies were calculated by first converting the average output activations for each condition into probabilities using Luce's choice axiom (Luce, 1959), which states that the probability of choosing one output is that output's activation divided by the sum of all output activations. These probabilities were then used to define a sub-interval for each location within the [0,1] interval. For instance, if calculated probabilities were red = 0.10, yellow = 0.70, blue = 0.05, and green = 0.15, the interval [0.0, 0.1[was attributed to the red location, [0.10, 0.80[to yellow, [0.80, 0.85[to blue, and [0.85, 1.0] to green. Finally, 21 random numbers between 0 and 1 were obtained, and each number falling within a location's sub-interval was counted as a search there. For example, the random value 0.23 would be counted as a search in yellow since it is in the sub-interval [0.10, 0.80[. Training ended when networks' observed frequencies of search were no longer significantly different from the expected frequencies.

Testing By analogy with the infant experiment, each network was tested in one experimental condition. Thus seven networks were tested in each of the 8 conditions (2 belief statuses x 2 belief locations x 2 search locations).

In test, we gave an input to the network (corresponding to a belief-induction trial in the infant experiment, i.e., first and second location of the object, and whether the actor watched the object move) and measured how far the generated output was from the target output.

Although networks were trained on four start and four end locations, they were tested only on 2 start and end locations, as in the infant experiment (Onishi & Baillargeon, 2005). The mean network error on these 2 nodes was calculated as the sum of squared difference between the network output and the target location, for each condition of the infant experiment.

For instance, in the infant study, 2 groups saw: object starts in green, moves to yellow, actor is not watching. One of the groups then saw the actor search in green (expected) and the other group saw her search in yellow (surprising). Networks were treated in the same way. The input: object starts in green, moves to yellow, actor is not watching, was given to 2 groups of networks. For one group we measured how far from "search in green" (target output activation: green = 1.0, yellow = 0.0) their output was and for the other group we measured the discrepancy from "search in yellow". How far the network was from predicting a search in green was calculated as the mean squared difference between its prediction and the actual location of search, as in Equation 1;

$$\left[(P_g - S_g)^2 + (P_y - S_y)^2\right] / 2 = \left[(1 - .85)^2 + (0 - .05)^2\right] / 2 = 0.0125$$

Equation 1

where P represents the model's prediction, S represents searching in the test trial, and the subscripts g and y represent the green and yellow boxes respectively. After training, networks from both groups output approximately 0.85 for green and 0.05 for the other nodes. In the "search in green" trial, the outputs for the green and yellow locations were compared to the target: green = 1.0, yellow = 0.0, resulting in an error of 0.0125 (as shown in Equation 1). The green and yellow target values were reversed for the "search in yellow" trial, yielding an error of approximately 0.8125. Thus, mean network error was greater for search in green than for search in yellow, quantifying the network's greater surprise.

To assess whether networks underwent developmental changes they were tested at the end of each output phase, that is, right before they recruited each hidden unit and once more at the end of training.

Results

Analyses of variance (ANOVAs) were performed after each output phase and at the end of training, with the factors belief status (whether the actor's belief was true or false), belief location (where the actor believed the object was), and search location (whether search was in the yellow or green box).

If networks had expectations consistent with a representational ToM in both true and false beliefs trials, we predicted a significant interaction between belief-location and search location and further, that this interaction would also be significant within each level of the belief factor (true and false). Further, we predicted that planned comparisons between the levels of search location would show lower error when search was in the location where the actor last

saw the object (as in Equation 1) at every level of the belief location by belief status interaction (TB-green, TB-yellow, FB-green, FB-yellow).

If however, networks had expectations consistent with an omniscient ToM, we predicted lower error for a search in the location of the object than for a search in the other location.

With 0 hidden units, networks displayed omniscient expectations about searching, as shown in Figure 3. Networks showed less error when search was in the location of the object. The belief location by search location interaction was significant overall, $F(1,48) = 8380$, $p < 0.001$ as well as within each belief status level, $Fs(1,24) > 1575$, $ps < 0.001$. Planned comparisons indicated that networks had lower error when search was in the object's end location for all levels of the belief location by belief status interaction, $Fs(1,12) > 512$, $ps < 0.001$.

With 0 hidden units, networks did not learn the probability distributions of the training patterns. The observed frequencies were significantly different from the expected frequencies in each of the eight conditions, $\chi^2s(3, N = 21) > 22$, $ps < 0.05$.

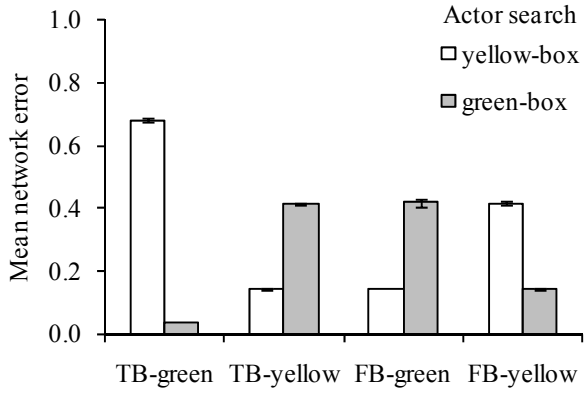

Figure 3: Mean network error and SE bars when networks had no hidden units. Networks' expectations were consistent with an omniscient ToM – error was lower when search was in the end location of the object.[1]

With one and two hidden units, networks' expectations continued to follow a pattern consistent with an omniscient ToM. However, with three hidden units, networks showed less error when search was in the belief location, i.e., the location in which the actor last saw the object. The belief location by search location interaction was significant overall, $F(1,48) = 215$, $p < 0.001$, as well as within each belief status level, $Fs(1,24) > 22$, $ps < 0.001$ Planned comparisons indicated that networks had lower error when search was in the last believed location of the object for all levels of the belief location by belief status interaction $Fs(1,12) > 8$, $ps < 0.02$.

Further, with three hidden units networks did learn the probability distributions during training. The observed frequencies were not significantly different from the expected frequencies for all conditions, $\chi^2s(3, N = 21) < .03$, $ps > 0.99$.

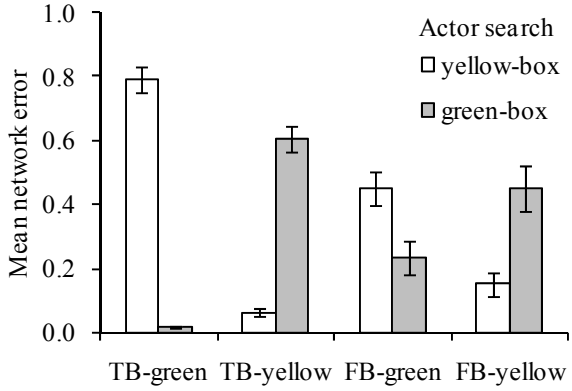

Figure 4: Mean network error and SE bars when networks had three hidden units. Networks' expectations were consistent with a representational ToM – error was lower when search was in the last believed location of the object.

Discussion

Our model went through the same developmental transition observed with older children tested with the explicit, standard false belief task. With insufficient hidden units, networks had expectations consistent with an omniscient ToM, but with three hidden units, networks had expectations consistent with a representational ToM. Further, with three hidden units, networks successfully learned the probability distributions present in training. Therefore, our model suggests that infants' expectations might also go through a similar transition. To obtain that transition, more true- than false-belief observations had to be present in training, lending computational support to the idea that children's (and infants') expectations might go through this transition because they develop the ability to override a default expectation of true belief (Leslie et al., 2005). We have shown that this default could be a natural result of experiencing more true- than false-belief behavior.

Comparison with Other Models

By contrast with other computational models of false belief tasks, our model is a developmental model that learns to solve the task without parameter manipulation.

O'Laughlin and Thagard (2000) and Triona et al. (2002) both covered the omniscient to representational developmental transition by parameter manipulation. O'Laughlin and Thagard (2000) manipulated the excitatory and inhibitory nature of the connections between the nodes, whereas Triona et al. (2002) directly manipulated the

[1] The greater difference between search locations in TB-green is an artefact from the infant experiment, in which the object did not move in TB-green (starting and ending in green), but did move in TB-yellow (starting in green and ending in yellow). This effect disappeared when TB-yellow was implemented in the model with the object starting and ending in yellow. In networks, a moving object is more difficult to process than a stationary object.

probability that the model would solve the task. Our model learns the implicit task and goes through the same developmental changes that children go through without direct parameter manipulation.

Goodman et al. (2006) showed that a Bayesian network having expectations corresponding to a representational ToM was more computationally complex than another Bayesian network having omniscient expectations. However, their model was restricted by the fact that the networks' architectures were designed by the experimenters and by the limited number of networks they implemented, which in turn limited the number of possible transitions the model could perform. Our model showed, in a single unified developmental system, that representational ToM expectations do require more computational power; networks required no hidden units to have expectations consistent with an omniscient ToM about others' searching, but required three hidden units to learn to have expectations consistent with a representational ToM. Both omniscient and representational ToM expectations were developed autonomously by the model, the later one building on the earlier one.

Summary

In sum, computational models are useful for providing insights into psychological phenomena and can lead to novel predictions or directions for experimental tasks. Our model provides a novel computational insight as to why children go through a developmental transition on the standard false belief task and predicts that infants' expectations might go through a similar transition on the implicit task.

Future directions could include an *avoidance* implicit FB task, in which the actor wants to avoid a noxious object, instead of search for an attractive object. Indeed, only after four years do children pass explicit FB tasks with avoidance goals, and expect an actor to correctly avoid a location when she has a false belief about its content. However, no avoidance implicit FB task has been used with infants. Would our model succeed an avoidance version of the implicit task only after it succeeded the approach task? Computational models such as ours are useful to understand how cognition undergoes developmental transitions.

Acknowledgments

This research was supported by a scholarship to V.G.B. from the Fonds Québécois de la Recherche sur la Nature et les Technologies, a grant to K.H.O. from the Fonds Québécois de la Recherche sur la Société et la Culture, as well as a scholarship to V.G.B. and grants to K.H.O. and T.R.S. from the Natural Sciences and Engineering Research Council of Canada.

References

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111*(4), 1036-1060.

Baluja, S., & Fahlman, S. E. (1994). *Reducing network depth in the cascade-correlation* (Technical report No. CMU-CS-94-209). Pittsburgh: Carnegie Mellon University.

Dennett, D. C. (1978). Beliefs about beliefs. *Behavioral and Brain Sciences, 1*(4), 568-570.

Goodman, N. D., Baker, C. L., Bonawitz, E. B., Mansinghka, V. K., Gopnik, A., Wellman, H., et al. (2006). Intuitive theories of mind: A rational approach to false belief. In *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society* (pp. 1382-1387). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Leslie, A. M., German, T. P., & Polizzi, P. (2005). Belief-desire reasoning as a process of selection. *Cognitive Psychology, 50*(1), 45-85.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.

O'Laughlin, C., & Thagard, P. (2000). Autism and coherence: A computational model. *Mind & Language, 15*(4), 375-392.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*, 255-258.

Perner, J., Leekam, S. R., & Wimmer, H. (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology, 5*(2), 125-137.

Shultz, T. R., & Bale, A. C. (2001). Neural network simulation of infant familiarization to artificial sentences: Rule-like behavior without explicit rules and variables. *Infancy, 2*(4), 501-536.

Shultz, T. R., & Bale, A. C. (2006). Neural networks discover a near-identity relation to distinguish simple syntactic forms. *Minds and Machines, 16*, 107-139.

Triona, L. M., Masnick, A. M., & Morris, B. J. (2002). What does it take to pass the false belief task? An ACT-R model. In *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society* (p. 1045). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72*(3), 655-684.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*(1), 103-128.