In Proceedings of ICCM - 2007- Eighth International Conference on Cognitive Modeling. 169 - 174. Oxford, UK: Taylor & Francis/Psychology Press.

Attention and Association Explain the Emergence of Reasoning About False Beliefs in Young Children

Paul Bello (Paul.Bello@rl.af.mil) & Perrin Bignoli (Perrin.Bignoli@rl.af.mil)
Air Force Research Laboratory
Information Directorate; 525 Brooks Rd.
Rome, NY 13441 USA
Nicholas Cassimatis (cassin@rpi.edu)
Rensselaer Polytechnic Institute
Department of Cognitive Science; 110 8th St.
Troy, NY 12180 USA

Abstract

Charting and explaining the development of young children's capacity to reason about mental states is a mainstay activity among developmental psychologists interested in how theory-of-mind (ToM) is acquired. These explanations are typically couched within one of the traditional frameworks for studying mental-state attribution: the *theory theory* and the *simulation theory*. This paper presents an analysis of the positions adopted on the issue of ToM development when subscribing to each of these frameworks, and argues for an alternative explanation of development based on a simple associative learning mechanism that appropriately shifts the child's cognitive focus of attention when asked to make predictions about the actions of others. We develop this notion within the confines of Wimmer and Perner's classic false belief task, and describe a cognitive model implemented within the *Polyscheme* computational cognitive architecture that realizes the development process.

Introduction

One of the most contentious areas in the literature on cognitive development concerns the acquisition of a fully matured theory-of-mind. Theory-of-mind (ToM) is broadly defined as the capacity to predict and explain the behavior of other agents by appealing to unobservable mental states such as their beliefs, desires, and intentions. Many byproducts of higher-order cognition such as the ability to participate fruitfully in discourse with others, our ability to assign blame through ascriptions of intentionality, to our very notion of ourselves as beings who have mental states that change through time seem to be served by our possession of a mature theory-of-mind. However, it remains unclear how such a remarkable ability emerges during the developmental process. An answer to this question clearly depends on the answers to a number of other more fundamental questions concerning the nature of mental states and how they are used within the human cognitive architecture. Nonetheless, various theories have given accounts of how ToM might develop, conditioned on relatively imprecise definitions of what it means for one to have a belief, desire, or an intention. We will give an alternative theoretical account of what it means to "believe" something (or believe that someone else believes something), and show how such an account can explain ToM emergence while avoiding the pitfalls encountered by the most common theoretical frameworks for explaining development. We will employ the Polyscheme computational cognitive architecture to explain shifts in children's responses on the false belief task (Wimmer & Perner 1983), which is widely regarded as a "gold standard" for illustrating differences between children who have and have not developed a mature ToM.

Frameworks for Explaining Development

The vast majority of researchers concerned with ToM development adopt one of three popular theoretical positions. The first and most widely adopted position is that of the "child as scientist." This position is usually called the *theory theory* (Carey 1985, Gopnik & Meltzoff 1997), and states that our knowledge of mental states is arranged in a theory-like set of interrelated concepts that subserve prediction and explanation of behavior. These "commonsense theories" comprise our domain-specific, intuitive notions about such domains as biology (Medin & Atran 1999), physics (Spelke et. al. 1992), and psychology. The second of these frameworks is the *simulation theory* and it's variants (Goldman 2006, Gordon 1986) which casts our ability to predict and explain behavior as the result of mental simulations within which we assume the perspective of the agent to be reasoned about, and use our own cognitive capacities to approximate those of the agent. We turn to the discussion of each of these positions with specific emphasis on the issue of development – the learning and/or maturational components of ToM. We do so by explaining the behavior of subjects in the classic false belief task. The task consists in showing subjects a storyboard comprised of the following pictorial representations, with appropriate narration by the experimenter:

- A little boy (Maxi) is in the kitchen with his mother, and he puts his chocolate on the counter.
- Maxi goes outside to play ball.
- While he is outside, Maxi's mother puts his chocolate in the kitchen cabinet.

The subjects are then asked where Maxi will look for his chocolate when he goes back into the kitchen. Many variations on this experiment exhibit the same general trend in the data: that between three and four years of age, children switch in the type of responses that they give from answering "the cabinet" (which is the typical three year old response) to answering with "the

counter" (which is the typical four year old response) (Wellman, Cross & Watson 2001). What does modern developmental theory say about this interesting pattern of responses?

Theory theory

One way to explain the distinction between three and four-year old behavior on the false belief task is to assume that children possess a set of interrelated concepts like belief, desire, and intention, which are used in predicting the actions that others will take. In general, to succeed on the task, the theory-theorist claims that children are able to think about the following:

> He/She believes that p.
>
> or:
>
> Persons who want that p and believe that q would be sufficient to bring about p and have no conflicting wants or preferred strategies will try to bring it about that q.

Where p and q are propositions. Being in possession of such knowledge requires an explicit concept of belief (as a predicate, for example), which requires a psychological theory within which to couch it. On the theory-theory account, development in children consists of gradually acquiring knowledge, and subsequently revising fragments (or the totality) of their theory. In order to use the theory, various information-processing mechanisms need to be properly functioning. So, according to the theory-theory, children's failure on the false belief task must either be the result of lacking the necessary body of concepts/knowledge or immaturity of the information-processing mechanisms that use this information in making predictions. The developmental process is some mixture of knowledge acquisition, a maturing facility for information-processing, and the acquisition of new concepts – specifically a so-called "representational" concept of belief. This raises some interesting questions: how can one separate contributions to failure made by limitations on information processing and those resulting from conceptual deficits? How is such a theory representationally structured, and what sort of implications might this have for learning? It seems that committing oneself to representing statements of the form: "He/She believes that p" requires us to relearn new theory for every agent in whom we come in contact with. How does the theory-theory account for the fact that targets often arrive at different conclusions and take different actions than we as predictors would? Does this imply that we need to have theory corresponding to each agent's inferential mechanisms? Finally, how could theory-theory explain the enormous number of beliefs that we attribute to agents with whom we've had no contact at all? We are certainly able to make detailed predictions about readers of this paper without ever having met them, and as it turns out, three year-old children are capable of doing the same (Nichols & Stich 2003).

Simulation Theory

Another way one could potentially predict and explain the behavior of others is through a process of mental simulation. The simulation theory was developed as a reaction to some of the problems that seem to be unavoidable when adopting a purely theory-driven framework for prediction. In order to succeed on the false belief task, an agent implementing simulation theory needs to be able to entertain beliefs of the form:

> p

Along with being able to think about such propositions, the agent must be able to imagine a counterfactual world in which it imaginatively identifies with the target it wishes to predict. Explaining development in light of this very general simulation theory is simple: children at three and four years of age do not possess different knowledge or conceptual structures, they merely become more adept at identifying with other people in the context of simulation. The imaginings necessary for simulation are ultimately linked to information processing capabilities, and changes in this capability produces observable developmental transitions in children. However, the simulation theory suffers from some immediate difficulties. Even if we are able to imaginatively identify with others and impute all of our own beliefs to them within the context of a counterfactual situation, it remains unclear how to account for differences between the simulator and the target. In the false belief task, it seems rather unsatisfying to assert that the subject "imaginatively identifies" with Maxi, and magically knows that the chocolate is on the counter. Rather, it seems as if the subject would need to employ a number of *overrides* to his current knowledge of the real world in order to make such a claim. The subject would also seemingly have to possess some form of knowledge about perceptual occlusion in order to come to an appropriate starting point for simulating Maxi's mental life while he is outside playing. While all of this extra knowledge may not be related to mental states, per se, it defeats the purpose of simulation – which is supposed to be an information-poor process as opposed to theory-laden, information rich-process.

Prior Modeling Work

The only prior modeling work on the developmental transition between three and four years of age is a Bayesian analysis presented in (Goodman et. al. 2006). The authors adopt a theory-theoretical perspective on development, and chart the transition of the child from a "copy theorist" (CT) who maintains that all events in the real world are copied as knowledge possessed by the target, and a so-called "perspective theorist" (PT) who uses other information to mediate what gets attributed to other agents. In the case of the false-belief task, this other information takes the form of knowledge about visual access. The learning process is surprise-driven, and is as follows:

1. Children start out with two theories, CT, and PT. CT is originally preferred to make predictions, since it

includes no extra information (in this case regarding the target's visual access).

2. Predictions are made with CT. Normally these predictions are successful and unsurprising since the target is usually in the presence of the predictor-agent, and there is no discrepancy in visual access.

3. Situations in which the target possesses a false belief are incorrectly predicted by CT, resulting in a high surprise value.

4. These same situations (and typical non false-belief situations) are correctly predicted by PT. Slowly, PT becomes the preferred device for predicting and explaining the target's behavior.

This is an elegant, rational explanation for development, but it leaves a few lingering questions to be answered. The models to be selected from are presumed to be probabilistic graphical models (Pearl 1988, 2000). In these models, there are nodes representing the "belief" of the target. It seems odd that on the one hand, theory-theory claims that a conceptual change happens between three and four regarding children's understanding of belief, yet these models presume children already have a sufficient representation of belief as copy-theorists. Rather, it is understanding that visual access is coupled to belief that allows children to become perspective-theorists. Much research has gone into children's understanding of visual access as it relates to beliefs, and the verdict seems to be in favor of a very early understanding that visual access factors into what others know (Lempers et. al. 1977, Gopnik & Slaughter 1991), leaving in question that switching between these two *particular* models should play any role in explaining four year old behavior. It's also clear, given a purely Bayesian interpretation of commonsense theories, that representing beliefs about beliefs about beliefs, *et cetera* becomes prohibitively costly from a computational standpoint, since entire networks would have to be recopied in memory to keep the relevant conditional probabilities from having unwanted influence on one another.

The Cognitive Substrate Hypothesis

One of the hallmarks of human cognition is the range of situations it is capable of adapting itself to. Presumably, our evolutionary forebears did not need to fabricate microprocessors, sail yachts, formulate grand unified theories, or do the majority of other activities that we are capable of doing in our current day and age. Instead, they most likely needed to be able to reason more generally about the nature of objects in their physical environments, and be able to make better predictions about their behavior in order to maximize their chances for survival in what we assume to be extremely inhospitable conditions. The Cognitive Substrate Hypothesis states that a (relatively) small set of properly integrated, domain-general computational mechanisms can provide a mechanistic explanation for much, if not all of higher-order cognition. While many different suggestions could potentially be made in defining a cognitive substrate as we've described above, our particular selection of domain-general mechanisms are motivated by developmental studies of physical reasoning. In order to successfully exist in a dynamic physical environment one must proficiently reason about a core set of domains including but not limited to: time, space, parts/wholes, paths, instrumental desires, events, identity/similarity, situations/worlds, and causality (including learning causal contingencies). We see computational functionality in these domains as being absolutely critical to the survival of most (if not all) primates, but especially humans - however it might be implemented.

Polyscheme: A Substrate Implementation

We have chosen to conduct our exploration using the Polyscheme computational cognitive architecture (Cassimatis 2005), which was originally designed to integrate multiple computational mechanisms corresponding to aspects of higher and lower-level cognitive processes. Polyscheme consists of a number of specialists which maintain their own proprietary representations that communicate with one another during problem-solving through coordination via a cognitive focus of attention. The selection of this particular implementation of attention is motivated by the existence of processing interference in the Stroop effect (Stroop, 1935), which suggests that multiple mental processes operate simultaneously (word and color recognition, for example). Visual attention has also been demonstrated as an integrative mechanism for inputs from multiple sensory modalities (Triesman & Gelade, 1980). Polyscheme is based on the notion that just as the perceptual Stroop effect can be generalized to higher-level non-perceptual cognition, that integrative perceptual attention suggests the existence of a higher-level cognitive focus of attention that is the mind's principle integrative mechanism.

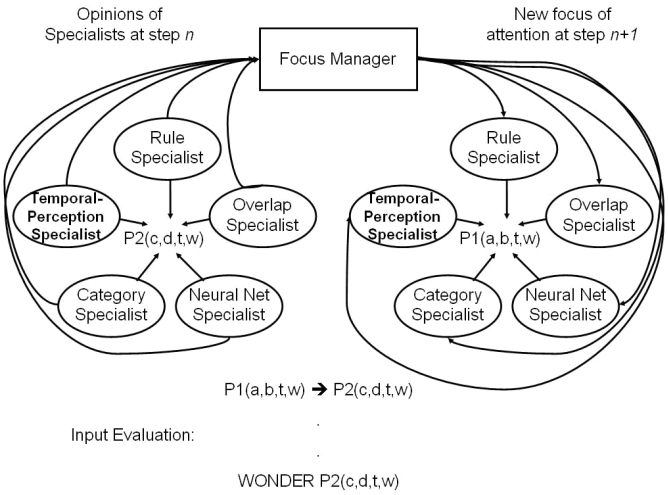

Figure 1: Focus Management in Subgoaling

Integration among specialists implementing their own computational methods is achieved through two basic

principles: the multiple implementation principle (MIP), and the common function principle (CFP). The common function principle states that each specialist implements a core set of common functions, including forward inference, subgoal generation, identity matching, and simulation of alternative worlds. The multiple implementation principle states that many different computational models (i.e. rules, neural networks, etc.) implement common functions. Details concerning how the focus of attention coordinates sequences of common functions generated by multiple representations can be found in (Cassimatis 2005).

Polyscheme: Some Formal Preliminaries

Strings of the form $P(x0, ..., xn, t, w)$ are called propositions. Simply stated, P is a relation (i.e. Loves, Hates, Color, MotherOf) over the set of objects xi during the temporal interval t (which could be a single time point, or a range of time points) in the world w which bears a truth value. A proposition's truth value is a tuple $< E+, E- >$ consisting of the positive (E+) and negative (E-) evidence for that proposition. Evidence takes on one of the following values: $\{C, L, l, m, ?\}$ representing certainly, very likely, likely, maybe, and unknown. So, if we are given the proposition $Likes(John, Mary, t, w)$, and come to find out its truth value is ¡m,m¿, we can say that at some time t, maybe John likes Mary, and maybe he doesn't. If at some later time say $t+1$, we find a note written by John to his friend, expressing his affection for Mary, we may update the truth value of this proposition, with $Likes(John, Mary, t + 1, w)$ taking on the value $< C, ? >$. Sometimes the letter E will appear in the proposition where a temporal argument would normally be. E represents "eternity" and denotes the temporal interval containing all other temporal intervals. Similarly, one might observe the letter R in the proposition where a world argument would normally be. R represents the real world, which consists of objects and relations in the environment appropriately transduced into propositions by Polyscheme's perceptual machinery. This is the world as Polyscheme experiences it. Letters other than R in the world argument of the proposition could represent hypothetical, future, counterfactual or past states of the world. We will exploit this functionality when describing how to perform ToM-driven inference.

A Substrate Mapping for Social Cognition

Based on a variety of evidence from the empirical literature, we have minimally extended the cognitive substrate for physical reasoning to accommodate reasoning about the mental states of others. The domain-general functionality for reasoning about physical objects includes a general purpose spatial competency, a mechanism for reasoning about identity, functionality that keeps track of the truth-values of perceived objects/relations through time, a simple associative learner, a system that keeps track of the truth value of propositions in different worlds (used for planning under uncertainty), and a rule-based reasoner. In addition, we add three new pieces of functionality that we believe are well-justified in the literature on child development: a mechanism to keep track of lines-of-sight (Hood et al. 1998), a mechanism for detecting (specifically) human agency (Guajardo & Woodward 2004), and a mechanism for generating exceptions for rules about self/other identity. Specifically, we use the latter to selectively override the version of *Leibnitz' Law*, which states that two objects are the same just in case they share all of the same properties. To even claim these three mechanisms as "additions" seems to be somewhat dubious as well, since we can imagine uses for all three of these functions in non-social cognition. For example, it is plausible to assume that an agency detector could be used to constrain search while performing object tracking, and we're convinced of the fact that overrides to the identity hypothesis are used frequently, especially in the case of early pretense, where features of a source object must be replaced with imaginary features of a target object. On our account, representing the "beliefs" of other agents consists of detecting agency, which causes the simulation of a counterfactual world w in which the identity $Same(self, other, E, w)$ is true. Beliefs held by "self" (i.e. propositions which are true in the real world) are *inherited* into the counterfactual world w with a slightly weakened truth value (in order to prevent immediate contradictions from arising). So, if at some time $t1$ self determines the location of the chocolate is on the counter, we have a corresponding proposition $Location(chocolate, counter, t1, w)$ which is true in w.

Reasoning About False Beliefs

As we have mentioned previously, the false belief task consists of the unexpected transfer of an object from one location to another that happens outside of the knowledge of a target agent, whose action toward this object is to be predicted by the subject. We claim that there is no gap in conceptual or theoretical knowledge differentiating three and four year old subjects. We claim that three year old subjects are in possession of all of the knowledge needed to pass the false belief task, but haven't yet learned to properly re-focus their attention on the target's line-of-sight. Our approach most closely resembles mental simulation, but also uses explicit knowledge in the form of rules to populate and guide the progress of simulation as it occurs. The set of rules that we assume both three and four year old children to be using consists of the following:

1. If an agent has line-of-sight on an object, then the agent knows the location of the object.
2. If an object is at a location it cannot be located anywhere else.
3. If an object is at a location at some time t, it will most probably be at that location at time t+1.
4. If an agent has line-of-sight on an object at time t, it will most likely have line-of-sight on that object at time t+1.
5. If an agent wants an object, and knows that the object is located at l, the agent will reach for the object at l.

We adopt a version of the "like me" hypothesis developed in (Meltzoff 2005), which broadly states that humans possess an innate faculty that posits equivalences between self and other. This idea has been supported through repeated observation of infants imitating the facial gestures of their care-givers even at forty minutes old. To do so, we use Polyscheme's identity predicate, $Same(x, y, E, w)$, in the context of a counterfactual world w, which serves as a mental simulation of x taking the perspective of y. When Polyscheme sees a proposition of the form: $Category(x, Agent, E, R)$, it immediately creates an alternative world w in which the proposition $Same(self, x, E, w)$ is true. We use an isomorphic version of the false belief task in which an agent named Sue sees a cookie in jar A, then goes outside, as in the classic task. While she is out, the cookie moves to jar B, and the subject (Polyscheme in this case) is asked where Sue thinks the cookie is.

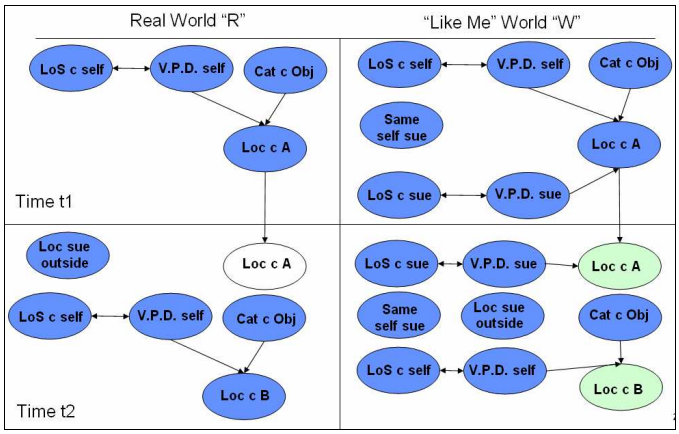

Figure 2: Inference in the False Belief Task

We will use figure 2 and figure 3 as visual references as we explain inference in the task. As we can see in 2, we represent propositions which are true in both the real world, and in the counterfactual world w. Upon noticing that Sue's category is *Agent*, w is created and seeded with the true proposition $Same(self, Sue, E, w)$. Self has a line of sight on the cookie at time t1, and thus it's location at jar A. In the counterfactual world in which Self is identical to Sue, Sue also has a line of sight on the cookie. Information about the cookie's location in the counterfactual world is inherited from information in the real world. So in w, the cookie's location is also at jar A at time t1. Since the cookie's location is at jar A at time t1, we infer that the cookie is likely at jar A at time t2. Similarly, self's line of sight at time t2 is initially on the cookie. The results of these inferences are also available in w. Now, Sue goes outside, but self stays inside and sees the cookie move from jar A to jar B. Now, self's line of sight is on the cookie at jar B, and the location of the cookie is now known by self to be at jar B. The interesting issue is that relations such as "location" are not indexed by agent names. The location of the cookie is just the location of the cookie. If this is the case, how do we separate self's knowledge of the location of the cookie from Sue's? Our speculation is that this is what separates three and four year old subjects.

Learning to Focus

How then, do four year old subjects successfully navigate the false belief task? We claim that four year old children selectively focus their attention on how information is acquired by the target in the simulation in order to make better predictions about how it will behave.

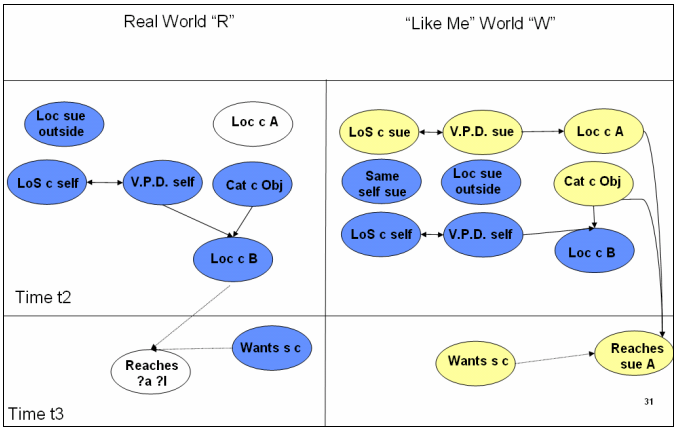

Figure 3: Re-focusing in the False Belief Task

In figure 3, we see Sue's line-of-sight highlighted. Even though self's line of sight still suggests that the cookie is in jar B, we re-focus on Sue's line of sight, and re-infer the location of the cookie to be at jar A. This re-focusing policy is a result of one of Polyscheme's specialists that monitors for conflicts in situations where self/other equivalences are drawn, and re-focuses it's cognitive focus of attention on other-specific information. In the false belief task, the self/other equivalence that causes our problem is that at time 2, and subsequently at time 3, there is a mismatch between self's line of sight and Sue's line of sight. One of Polyscheme's specialists detects this mismatch and re-focuses on Sue's line of sight, re-inferring the location of the cookie to still be at jar A. Learning what to focus on is the crucial linchpin in the developmental process. Polyscheme's associative learner keeps track of which propositions are true every time an action is either taken by self in the real world, or predicted in counterfactual worlds. The learning process is driven by bad predictions. Polyscheme learns to associate the appearance of certain propositions (such as line-of-sight) with potential contradictions. Once it has accumulated a prioritized list of these propositions, they are made available to Polyscheme's conflict-resolver. If the propositions in conflict have an agent-name other than "self" as an argument, the conflict-resolver re-focuses attention on the proposition containing the other agent's name. In the false belief task, Polyscheme makes a number a bad predictions about where Sue thinks the cookie is, and learns to associate line-of-sight with misprediction. The conflict resolver will then re-focus attention

on Sue's line-of-sight, which in the context of simulating Sue's mind, will produce the correct prediction.

Summary

We have shown that learning to keep track of situations in which there are discrepancies between line-of-sight in agents, and using these discrepancies to focus attention on the line-of-sight of the target is a plausible explanation for the emergence of facilitation on the false-belief task. This explanation avoids a number of the problematic corollaries of adopting a more classical stance on the ToM issue. Excessive duplication of propositions and rule-fragments is avoided through the simulation of counterfactual states of affairs which inherit directly from our experience of the real world. By an large, propositions are agent-independent, alleviating the need to re-tag pieces of knowledge as being associated to the various agents whom we wish to make predictions about. Difficulty in learning about mental states due to unobservability is avoided, since via the inheritance and the simulation process, we have access to these structures. We suspect that the late emergence of false belief in the third to fourth year is caused by the general lack of training examples in which we need to keep track of our own line of sight in relation to the line of sight of others. While some may claim that bouts of joint attention between infants and others constitute training examples where infants must keep track of their own line of sight in relation to the lines of sight of others, it's not clear how the specific task of action-prediction interacts with the mental accounting being performed in these cases. It might be that the added complexity delays successful performance on ToM tasks until sometime between the three and four year marks in the same way that learning past-tense information in language usage is delayed.

References

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

Carey, S. (1985). Conceptual change in childhood. MIT Press/Bradford Books, Cambridge, MA.

Gopnik, A. & Meltzoff, A.N. (1997). Words, thoughts, and theories. Cambridge, Mass. Bradford, MIT Press.

Medin, D.L. & Atran, S. (1999). Folkbiology. MIT Press.

Spelke, E.S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. Psychological Review, 99, 605-632.

Goldman, A. (2006). Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading. Oxford University Press.

Gordon, R. (1986). Folk Psychology as Simulation. Mind and Language 1, 158-171; reprinted in Davies, M. and Stone T., eds., 1995, Folk Psychology: The Theory of Mind Debate. Oxford: Blackwell Publishers.

Wellman, H.M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. Child Dev, 72(3):655684.

Nichols, S. & Stich, S. (2003). Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding of Other Minds. Oxford University Press.

Wellman, H.M. (1990). The child's theory of mind. Cambridge, MA: MIT Press.

Goodman, N.D., Bonawitz, E.B., Baker, C.L., Mansinghka, V.K., Gopnik, A., Wellman, H., Schulz, L. and Tenenbaum, J.B. (2006). Intuitive theories of mind: A rational approach to false belief. Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society.

Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference." Morgan Kaufmann.

Pearl, J. (2000). Causality. Cambridge.

Lempers, J.D., Flavell, E.R., & Flavell, J.H. (1977). The development in very young children of tacit knowledge concerning visual perception. Genetic Psychology Monographs 95: 353.

Gopnik, A. & Slaughter, V. (1991). Young children's understanding of changes in their mental states. Child Development, 62, 98-110.

Cassimatis, N.L. (2005). Integrating Cognitive Models Based on Different Computational Methods. In Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society.

Stroop, J.R. (1935). Studies of interference in serial verbal reactions. Journal of Experimental Psychology, 18, 622-643.

Treisman, A.M. & Gelade, G. (1980). A feature integration theory of attention. Cognitive Psychology, 12, 97-136.

Guajardo, J.J., & Woodward, A.L. (2004). Is agency skin-deep? Surface attributes influence infants sensitivity to goal-directed action, Infancy , 6, 361-384.

Hood, B., Willen, J., & Driver, J. (1998). Adults' eyes trigger shifts of visual attention in human infants. Psychological Science, 9 (2), 131-134.