

Intuitive Theories of Mind: A Rational Approach to False Belief

Noah D. Goodman¹, Chris L. Baker¹, Elizabeth Baraff Bonawitz¹, Vikash K. Mansinghka¹
Alison Gopnik², Henry Wellman³, Laura Schulz¹, Joshua B. Tenenbaum¹

¹Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology,

²University of California, Berkeley, ³University of Michigan

Abstract

We propose a causal Bayesian model of false belief reasoning in children. This model realizes theory of mind as the rational use of intuitive theories and supports causal prediction, explanation, and theory revision. The model undergoes an experience-driven false belief transition. We investigate the relationship between prediction, explanation, and surprise; this is used to interpret an empirical study of children’s explanations in an extension of the false belief task. Our study includes the standard outcome, surprising to younger children, and a novel “Psychic Sally” condition that challenges older children with an unexpected outcome.

In everyday life, humans constantly attribute unobservable mental states to one another, and use them to predict and explain each others’ actions. Indeed, reasoning about other people’s mental states, such as beliefs, desires, and emotions, is one of our main preoccupations, and the source of some of our most virtuosic inferences. These abilities have been collectively called theory of mind (Premack and Woodruff, 1978), and have become one of the most well-studied, and contentious, areas in modern psychology. In particular, much research has focused on the phenomenon of *false belief*: the ability to infer that others hold beliefs which differ from the (perceived) state of the world. An often used assay of this ability is the standard false belief task (Wimmer and Perner, 1983): the subject is read a story in which Sally places her toy in a cabinet, then goes out to play. In Sally’s absence her toy is moved to a basket (causing her belief to be false). The subject is asked to predict Sally’s action: “when Sally comes back in, where will she look for her toy?” Many authors have reported that performance on this task undergoes a developmental transition, from below-chance to above-chance performance, in the third or fourth year of life (see Wellman et al. (2001) for a review and meta-analysis, though see also Onishi and Baillargeon (2005)).

Another influential thread of research in cognitive science has supported the idea that human behavior is approximately rational within its natural context (Anderson, 1990). Within cognitive development both strong and weak versions of this thesis are possible. On the strong interpretation children respond and learn rationally throughout development; developmental stages can thus be analyzed as individually optimal, in context, and collectively as a rational progression driven by experience. On the weak reading it is only the final, mature,

state which can be expected to be rational. The contrast between these interpretations has played out vividly in the microcosm of research on false belief (cf. Leslie, 1994; Gopnik and Wellman, 1992).

It has also been suggested (Carey, 1985) that domain knowledge, such as theory of mind, takes the form of intuitive theories, or coherent “systems of interrelated concepts that generate predictions and explanations in particular domains of experience” (Murphy, 1993). This viewpoint leads to an interpretation of the false belief transition as a revision of the child’s intuitive theory from a *copy theorist* (CT) position about beliefs (ie. beliefs are always consistent with the world) to a *perspective theorist* (PT) position (ie. beliefs can be false). However, the false belief transition is slow: children do not immediately achieve false belief when exposed to evidence *prima facie* incompatible with the CT position (Amsterlaw & Wellman, In Press; Slaughter & Gopnik, 1996). This presents a puzzle for strong rationality: how could it be rational to maintain a CT position about beliefs in the face of prediction failures, and, if it is rational, why (and when) should this position be revised given additional counter-evidence?

In this paper we give a formal model of theory of mind as the rational use of intuitive theories. This account illuminates the revision puzzle and allows us to explore the relationship between prediction and explanation. We present only the apparatus necessary to illuminate the above puzzles in the case of the standard false belief task, leaving many important elaborations for future work. An intuitive theory supports several core competencies, including causal prediction, explanation, and revision in response to new evidence. Gopnik et al. (2004) have suggested that intuitive theories may be represented as causal Bayesian networks (Pearl, 2000). We introduce two Bayesian network models and show how they support prediction, explanation, and revision¹. We propose that these two models, which differ only in the dependence of Sally’s belief on her visual perspective, coarsely approximate the intuitive theories which generate the CT and PT positions.

We probe these ideas experimentally by investigating the coherence of children’s explanations with their

¹Our formal analysis takes place at Marr’s computational level of modeling (Marr, 1982), that is, we describe the competencies, but not the algorithms (ie. procedures), of cognition.

predictions in cases when these predictions succeed and when they fail. This is arranged via the false belief task with both possible outcomes: the standard outcome (surprising to CTs), and a “psychic” outcome (surprising to PTs).

Formal Models

In the standard false belief task, described earlier, the story begins with Sally seeing her toy in the cabinet. As the story continues there are only three (observable) variables that have multiple outcomes: the final position of the toy, Sally’s visual access to the final position (ie. whether the doors to the cabinet and box are open), and Sally’s action upon re-entering the room. Thus we have the variables *World*, *Visual Access*, and *Action* available to our models (see Table 1 for descriptions). In addition, there are two unobservable mental state variables: Sally’s belief about the location of her toy, *Belief*, and her *Desire*. We simplify the, presumably sophisticated, sub-theory of desire (see Baker et al., In Press), by collapsing desires into one variable, which indicates whether Sally’s primary desire is her toy. (Formally, we marginalize out all other variables in this sub-theory.)

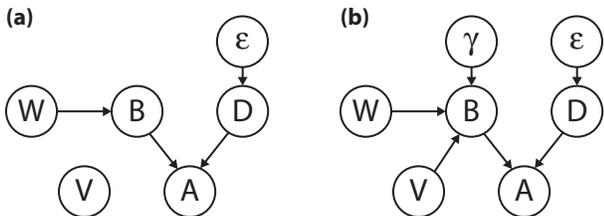


Figure 1: The dependency graphs of our Bayesian Network Models: (a) CT model, (b) PT model. Variables abbreviated by their first letter.

We specify the relationships between these variables by fixing their joint distribution, given by a causal Bayesian network. The pattern of conditional dependencies, given by the directed graphs in Fig. 1, codifies the standard view that action is determined by beliefs and desires², and that belief is affected by the state of the world. In the PT model belief is also dependent on access.

The conditional dependencies are parameterized by conditional probabilities in Table 1. The conditional probability table for action describes a simple case of the *rational agent assumption*: a person will act rationally, given her beliefs, to achieve her desires; in this case, if Sally wants her toy she will go to the location she believes it to be in, otherwise she goes to either location with equal probability (surely a simplification, but sufficient for present purposes). The variable *Desire* has prior probability $1 - \epsilon$, which will be large for desirable objects (such as a toy).

For the CT model, *Belief* is constrained to equal *World*. This is also true for the PT model when *Visual*

²We could equally have implemented a Desire-Action model for the CTs. This choice remains an interesting empirical issue.

Access is present, but without access Sally maintains her original belief, *Belief* = 0, with probability $1 - \gamma$. This parameter, γ , represents all the reasons, outside of the story, that Sally might change her mind: her sister might tell her the toy has moved, she may have E.S.P., she may forget that she actually left her toy in the cabinet....

We assume beta distribution priors on ϵ and γ . In the example simulations described below (Figures 3 and 2) the hyper-parameters were set to $\beta(1, 10)$ for ϵ , indicating that Sally probably wants her toy, and $\beta(1, 5)$ for γ , indicating that she is unlikely to change her belief (lacking access).

Prediction

Having represented our models as distributions, rational predictive use is now prescribed by the algebra of probabilities: conditioned on observation of some subset of the variables, a posterior distribution is determined that predicts values for the remaining variables. In addition, these models are *causal* theories: they support predictions of the outcome of interventions (via the causal *do* operator).

Take the example in which Sally’s toy has been moved, but Sally doesn’t have visual access to this new location (see schematic Fig. 3(a)). We may predict the probability that Sally will look in the Basket by marginalizing the unobserved variables and integrating out the parameters. The result is shown in Fig. 3(b); we see that the CT model “fails” the false belief test by predicting that Sally looks in the new (true) location, while the PT model “passes” by predicting the original location. For both models, however, even the unpredicted outcome is not impossible (as no probability in Fig. 3(b) is zero).

Theory Revision

Strong rationality requires an agent to rationally balance different available intuitive theories against each other, as well as to use each theory in a rational way. How should a rational theory-user combine, or select, possible theories of a domain, given the body of her experience? Fortunately, the algebra of Bayesian probability continues to prescribe rational use when there are competing models: the degree of belief in each model is given by its posterior probability given previous experience. We may then write down a belief weight comparing belief in the PT model to the CT model:

$$W_{PT/CT} = -\log(P(PT|X)/P(CT|X)),$$

where X represents experience in previous false belief settings.

When $W_{PT/CT}$ is strongly negative the contribution from PT is negligible, and the agent behaves as though it is a pure CT. If evidence accumulates and shifts $W_{PT/CT}$ to be strongly positive the agent behaves as a PT. In the liminal period, when $W_{PT/CT}$ is near zero, the agent maintains both theories. In this case the behavior of the agent may exhibit very sensitive dependence on new evidence: as $W_{PT/CT}$ fluctuates in response to immediate evidence the agent’s behavior may switch between the two models. For instance, if $W_{PT/CT}$ is slightly negative

Variable	Description	States
<i>World</i> (W)	Location of the toy.	0: Original location, 1: New location.
<i>Access</i> (V)	Could Sally see the toy moved?	0: No, 1: Yes.
<i>Action</i> (A)	Where Sally looks for her toy.	0: Original location, 1: New location.
<i>Belief</i> (B)	Where Sally thinks the toy is.	0: Original location, 1: New location.
<i>Desire</i> (D)	Sally’s primary desire.	1: To find the toy, 0: Anything else.

$P(A = 1 B, D)$	B	D
0	0	1
1	1	1
0.5	0	0
0.5	1	0

$P_{PT}(B = 1 W, V)$	W	V
0	0	1
1	1	1
γ	0	0
γ	1	0

$P_{CT}(B = 1 W)$	W
0	0
1	1

Table 1: The random variables and probability distribution tables for our models.

the CT model is initially applied, if evidence then surfaces which favors PT over CT the agent may begin to use the PT model; the reverse (PT to CT fluctuations) should also be possible.

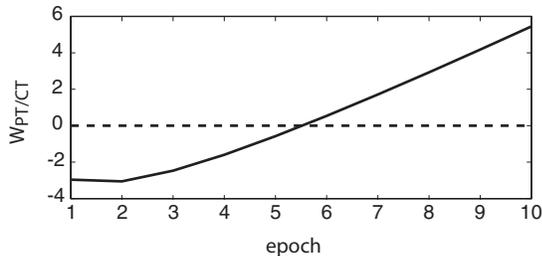


Figure 2: The log-posterior odds ratio over data epochs, showing the false belief transition from CT to PT. (Parameters integrated numerically by grid approximation.)

In Fig. 2 we plot $W_{PT/CT}$ evaluated on accumulating “epochs” of experience. Each epoch consists of trials with (W, V, A) observed, but (D, B) unobserved. The trials in each (identical) epoch encode the assumptions that visual access is usually available, and that, in instances without access, the protagonist often has a correct belief anyway (eg. to a child, his parents often appear to have preternatural knowledge). (Specifically, each epoch is twenty $(W=1, V=1, A=1)$ trials, six $(W=1, V=0, A=1)$ trials, and one $(W=1, V=0, A=0)$ trial.) The expected transition from CT to PT does occur under these assumptions. This rational revision depends on the particular character and statistics of experience, thus a developmental account is incomplete without empirical research on the evidence available to children in everyday life.

How can we understand the delayed confirmation of the PT model? First, in the initial epoch, the CT model is preferred due to the Bayesian Occam’s razor effect (Jeffreys and Berger, 1992): the PT model has additional complexity (the free parameter γ), which is penalized via the posterior probability. However, the data itself is more likely under the PT model – because some of the data represent genuine false belief situations. As

data accumulates the weight of this explanatory advantage eventually overcomes complexity and the PT model becomes favored.

It has been suggested (eg. Bartsch and Wellman, 1989) that explanations are a useful probe into the liminal period of theory revision, so we turn next to a discussion of this competency.

Explanation

An important function of intuitive theories is to explain observations by hypothesizing states of unobserved variables, from which the observations follow. To model this explanation competency we first recast our models into a deterministic *explicit-noise form*, by introducing additional explanatory variables, in order that observations will follow necessarily from unobserved variables. Explanation can then be described as inference of a *complete explanation* – a setting of all variables – and communication of a portion of this complete explanation, the *explanans*. A partial account of the explanans can be given by appealing to the *principle of surprise: a good explanans will address all of the ways in which an explanation is surprising*. (See Halpern and Pearl (2001) for a related approach.)

Formal Description Our PT model becomes deterministic if we introduce an *External Information* variable, E_γ (with prior probability γ), to explicitly represent events which cause changes in belief (in the absence of access). (For completeness, an additional variable that determines the object of alternate desires could be included; for technical reasons this variable has minimal effect on explanation, and has been omitted for clarity.)

For a Bayesian model, cast in explicit noise form, explanatory inference is dictated by the posterior distribution, conditioned on observations: the degree of belief in each complete explanation is given by its posterior probability (eg. Fig. 3(c)).

Which explanans will be given to explain an observation? Certainly it must be a partial report of some acceptable complete explanation, and the principle of surprise suggests a further criterion that must be satisfied. To implement this idea, surprise may be formalized

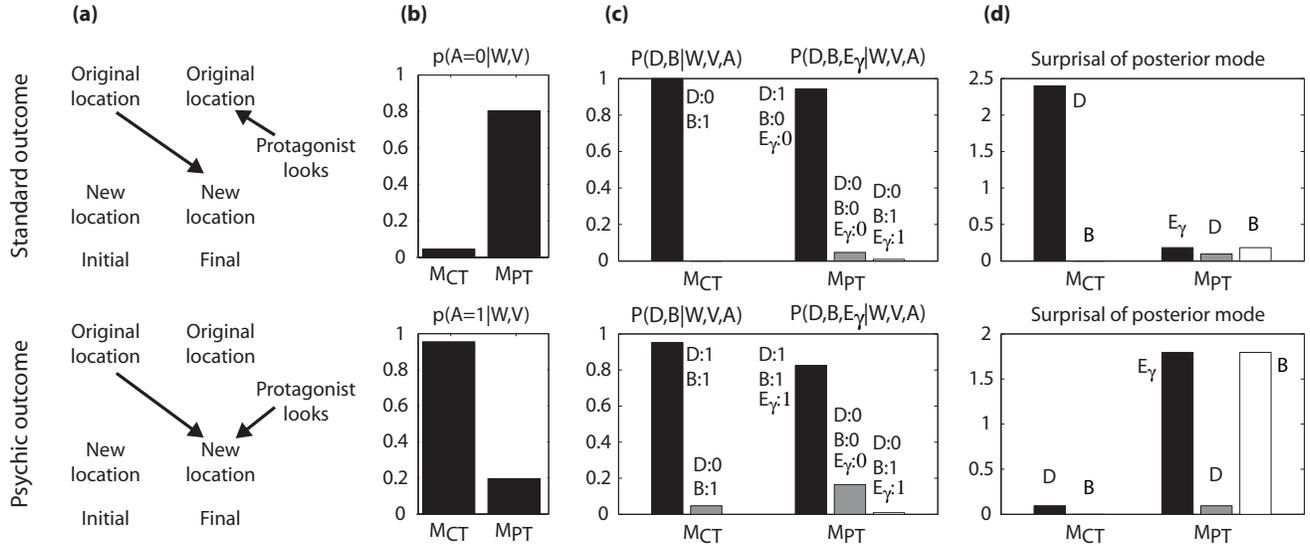


Figure 3: Comparing the Models: (a) Example situations, in both cases $W=1$, $V=0$, for the Standard outcome $A=0$, for the Psychic outcome $A=1$. (b) The predicted probability of each outcome. (c) Posterior probability of configurations of hidden variables, after observing the outcome. This indicates degree of belief in the corresponding complete explanation. (d) Surprise value of each variable in the modal configuration.

by the information-theoretic surprisal of a variable setting with respect to a reference distribution. For variable setting $Y=y$, given other settings $\{X_i=x_i\}$ already communicated, the surprisal is:

$$S(Y=y|\{X_i=x_i\}) = -\log_2(P(Y=y|\{X_i=x_i\}, Obs)).$$

Where Obs represents the observations common to reference and explainer: in our case W , V (and the model). (This may be derived by assuming the explanans is given to another agent with the same model, lacking only the final observation (the action taken), hence using the predictive posterior as reference distribution.)

Applying the principle of surprise, we may now give criteria for an explanation to be consistent with a model: 1) there must be a complete explanation, extending the explanans, with high posterior under the model, and 2) the state of no variable in this complete explanation should have high surprisal given the explanans. These criteria are qualitative: “high” is interpreted via an arbitrary threshold.

Applying the Consistency Criteria To analyze the example situations we set the thresholds to posterior probability >0.5 , and surprisal >1 . (We used the same thresholds for the experimental analysis below.) Only the highest probability complete explanations, for each model and condition, remain after thresholding (see Fig. 3(c)). For the PT model these complete explanations are ($D=1, B=0, E_\gamma=0, A=0, W=1, V=0$) in the Standard outcome (eg. informally, “Sally wanted to find her toy, thought it was in the original location,...”), and ($D=1, B=1, E_\gamma=1, A=1, W=1, V=0$) in the Psychic outcome (eg. “Sally wanted her toy, thought it was in the new location, someone had told her so,...”). For the CT model the complete explanations are ($D=0, B=1, A=0, W=1, V=0$) in the Standard

outcome (eg. “Sally wanted something else,...”), and ($D=1, B=1, A=1, W=1, V=0$) in the Psychic outcome (eg. “Sally wanted her toy, thought it was in the new location,...”).

Surprisals in these complete explanations are shown in Fig. 3(d) – it is only in the unexpected outcome cases that any variable is very surprising. However, in the unexpected outcome case of the PT model B no longer has high surprisal given the value of E_γ (ie. $S(B=1) > 1$, while $S(B=1|E_\gamma=1) \ll 1$), and vice versa. Now applying the second consistency criterion, only explanations containing the assertion $D=0$ are consistent with the CT model for the Standard outcome; for the Psychic outcome an explanation must mention $E_\gamma=1$ or $B=1$ to be consistent with the PT model. (Note that the explanation “Sally thought her toy was in the new location” ($B=1$) is consistent with both models, for the Psychic outcome – explanations in this condition are thus slightly less discriminable by our criteria.)

Children’s Explanations

If children are using intuitive theories of mind, as described here, they should exhibit coherence between prediction and explanation, possibly modulated by immediate experience in the case of liminal children – those on the threshold of the false belief transition. Accordingly, we investigated the different explanations that PT-predicting and CT-predicting children generated when presented with the two possible outcomes to the standard false belief story.

Participants Eighteen children (R=3;1-4;9, M=3;11) were tested in a quiet corner of an interactive science exhibit at a local museum. Parents were visible to the child, but could not hear the details of the story and were instructed not to interact with the child during the

study.

Materials Three picture books were created. The first book presented a guessing game story unrelated to the false-belief task. The second two books followed the standard Sally-Anne style narrative, with cartoon pictures depicting the events throughout the book. One book ended with the main character searching for her toy in the original location (Standard outcome – consistent with PT predictions); the other book ended with the main character searching for his toy in the new location (“Psychic” outcome – consistent with the CT prediction). These stories are equivalent to the situations of Fig. 3(a).

Procedure All children first saw the unrelated story, which simply familiarized the child with the experimenter and with generating guesses. The order of the remaining books was randomly chosen.

Standard outcome book: Sally was shown hiding her teddy-bear in a basket before going outside. Children were asked to point-out where the teddy-bear was being hidden. Then a mischievous character, Alex, was shown moving the teddy-bear from the basket to the box while Sally was away. Children were asked where the teddy-bear was moved. On the third page, Sally starts to come back into the room, and the children were asked, “Here comes Sally. Where do you think Sally is first going to go to get her toy?” After the children responded, the next scene depicted Sally going to the basket to get her toy. The children were then prompted with, “Sally went to look for her bear in the basket. Sally’s bear is really in the *box*. But Sally is looking for it in the *basket*! Why is she looking there?” The children were given the chance to respond. If they were unable to provide an explanation or provided uninformative information, the experimenter repeated, “Yes, but she’s looking for it way over here. What happened?” If the child was still unable to generate a response, a final forced choice question was given, “Suppose one of these two things happened. Do you think Sally is looking over here because she wants to, or do you think Sally is looking over here because she didn’t see it moved and it’s really over here,” (pointing to box). The order of these options was randomized.

Psychic outcome book: The procedure and dialog were essentially identical to the Standard outcome book, except the main character searched for his missing item in the location to which the item was moved, not where it was left. There were also superficial differences involving different characters (Billy & Anne), different objects (a cookie), and different locations (a drawer and a cabinet). Predictions and explanations were elicited as before.

Results and Discussion Children were coded as PT-predictors or CT-predictors according to the responses they gave to the prediction question in each book. Eight children selected the new location in both conditions and were coded as CT-predictors, 8 children selected the original location in both conditions and were coded as PT-predictors, and 2 children were coded as PT-predictors in one condition and CT-predictors in the other. There were no order effects. Responses to the explanation ques-

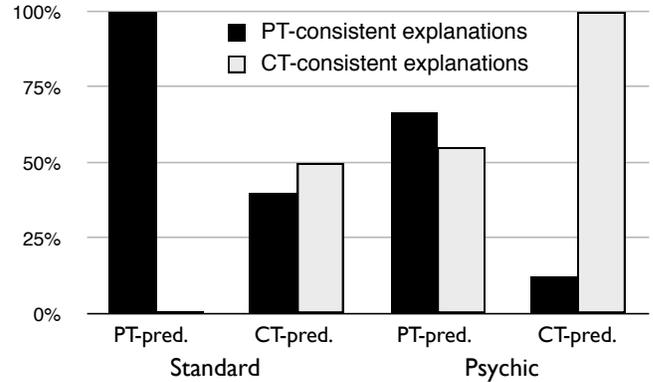


Figure 4: Portion of children’s explanations consistent with each model, broken down by prediction-group and outcome condition.

tions were scored by two coders, one who was blind to the group type for each child and to the formal model; a single inconsistency was resolved by discussion.

Explanations were coded for the mention and value, if any, of each variable. (The variables were as in Table 1, supplemented with External Information (E_γ) and Initial World variables.) For example, one (PT-predictor) child explained the Psychic outcome by “I think he heard his sister going over there,” and this was coded as $E_\gamma=1$. Another (CT-predictor) child explained the Standard outcome by “well, that’s where she wants to look,” which was coded $D=0$.

The consistency criteria described earlier were then applied to the explanations: each explanation was coded consistent or inconsistent with each model. (See the above section, Applying the Consistency Criteria, for a discussion of which explanations are formally consistent with each model.) Three explanations were consistent with both models, one explanation was consistent with neither model, and one explanation was un-resolvable within this framework and was therefore excluded.

The data (Fig. 4) suggest that prediction was related to explanation in the manner required by our models: prediction-type had a strong effect on the explanations generated in each book. Indeed, collapsing across outcome conditions, PT-predictors offered PT-consistent explanations significantly more than CT-predictors, and CT-predictors offered CT-consistent explanations significantly more than PT-predictors ($p < 0.01$ by Fisher exact tests). Looking at the individual conditions, we see that unsurprising outcomes (Standard for PT-predictors, Psychic for CT-predictors) led to almost uniform agreement between prediction and explanation. In contrast, for surprising outcomes (Psychic for PT-predictors, Standard for CT-predictors) children were almost equally likely to give explanations consistent with either model.

There are several possible explanations for children’s mixed responses in the surprising conditions. It is possible that children’s fragile explanatory capabilities were simply over-taxed when presented with prediction failures. Another possibility is that many of the children

were sufficiently close to the false belief transition that their theories should be considered mixtures of competing models, rather than purely CT or PT (a possibility reminiscent of Siegler (1996)). In this case, a sensitive dependence of belief weight on immediate evidence could cause children to resort to the competing model when confronted with failures of their primary model, as described above. That is, these liminal children are faced with a choice: stick with the theory they believe more strongly, or use a less likely theory with greater explanatory ability. It is intriguing to consider how a more complete account of fluctuating models in liminal theory use, perhaps via hierarchical Bayesian modeling, could give a rational basis for these phenomena. Clearly, further experiments and modeling work will be necessary to clarify these suggestions.

Conclusion and Future Directions

The history of developmental psychology has been filled with latent tension between the view that children are incomplete minds biding their time until full maturation, and the view that they are rational agents bootstrapping their way to an understanding of the world. The notion that children are strongly rational is alluring, as it would provide a uniform principle from which to understand development.

We have outlined a computational account of theory of mind as applied to the false belief task. This framework realizes theory of mind as rational use and revision of intuitive theory. Few formal models have been previously presented to account for false belief, and, to the best of our knowledge, none of these other models gives a strongly rational account. The CRIBB model of Wahl and Spada (2000), for instance, approaches failures of false belief as the result of limited processing capability.

This computational account sheds some light on the puzzle of rational theory revision, and suggests that a great deal may be learned by considering the coherence of children's predictions and explanations. Our prediction-explanation experiment included a novel Psychic outcome condition designed to be surprising to children who passed the false belief test. Though preliminary, our results suggest an overall coherence of explanation with prediction, but also point to the need for further study of the explanations of surprising outcomes given by liminal children.

The present account of the false belief transition as rational theory change is incomplete in important ways. After all, our agent had only to choose the best of two known models. This begs an understanding of the dynamics of rational revision near threshold and when the space of possible models is far larger. Further, a single formal model ought ultimately to be applicable to many false belief tasks, and to reasoning about mental states more generally. Several components seem necessary to extend a particular theory of mind into such a framework theory: a richer representation for the propositional content and attitudes in these tasks, extension of the implicit quantifier over trials to one over situations and people, and a broader view of the probability distri-

butions relating mental state variables. Each of these is an important direction for future research.

Acknowledgments

Thanks to the Boston Museum of Science, participants of the McDonnell Workshops (2005), R. Saxe, T. Lombrozo, and T. Kushnir. This research was supported by the McDonnell Foundation Causal Learning Collaborative Initiative.

References

- Amsterlaw, J. and Wellman, H. (in press). Theories of mind in transition: A microgenetic study of the development of false belief understanding. *Journal of Cognition and Development*.
- Anderson, J. R. (1990). *The adaptive character of thought*. Erlbaum, Hillsdale, NJ.
- Baker, C. L., Tenenbaum, J. B., and Saxe, R. R. (in press). Bayesian Models of Human Action Understanding. *Advances in Neural Information Processing Systems 18*.
- Bartsch, K. and Wellman, H. (1989). Young children's attribution of action to beliefs and desires. *Child Dev*, 60(4):946–964.
- Carey, S. (1985). *Conceptual change in childhood*. MIT Press/Bradford Books, Cambridge, MA.
- Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., and Danks, D. (2004). A theory of causal learning in children: causal maps and Bayes nets. *Psychol Rev*, 111(1):3–32.
- Gopnik, A. and Wellman, H. (1992). Why the child's theory of mind really is a theory. *Mind and Language*, 7:145–171.
- Halpern, J. Y. and Pearl, J. (2001). Causes and explanations: a structural-model approach. Part II: Explanations. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*.
- Jeffreys, W. and Berger, J. (1992). Ockham's Razor and Bayesian Analysis. *American Scientist*, 80:64–72.
- Leslie, A. M. (1994). Pretending and believing: issues in the theory of ToMM. *Cognition*, 50(1-3):211–238.
- Marr, D. (1982). *Vision*. Freeman Publishers.
- Murphy, G. L. (1993). Theories and concept formation. In Mechelen, I. V., Hampton, J., Michalski, R., and Theuns, P., editors, *Categories and concepts: Theoretical views and inductive data analysis*. Academic Press.
- Onishi, K. H. and Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, 308(5719):255–258.
- Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge University Press.
- Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *The Behavioral and Brain Sciences*, 4:515–526.
- Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking*. Oxford University Press.
- Slaughter, V. and Gopnik, A. (1996). Conceptual coherence in the child's theory of mind: training children to understand belief. *Child Dev*, 67(6):2967–2988.
- Wahl, S. and Spada, H. (2000). Children's reasoning about intentions, beliefs and behavior. *Cognitive Science Quarterly*, 1(1):3–32.
- Wellman, H. M., Cross, D., and Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. *Child Dev*, 72(3):655–684.
- Wimmer, H. and 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.