Intuitive Theories of Mind: A Rational Approach to False Belief

Noah D. Goodman[1], Chris L. Baker[1], Elizabeth Baraff Bonawitz[1], Vikash K. Mansinghka[1]
Alison Gopnik[2], Henry Wellman[3], Laura Schulz[1], Joshua B. Tenenbaum[1]
[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology,
[2]University of California, Berkeley, [3]University of Michigan

Abstract

We propose a rational analysis of children's false belief reasoning. Our analysis realizes a continuous, evidence-driven transition between two causal Bayesian models of false belief. Both models support prediction and explanation; however, one model is less complex while the other has greater explanatory resources. Because of this explanatory asymmetry, unexpected outcomes weigh more heavily against the simpler model. We test this account empirically by showing children the standard outcome of the false belief task and a novel "psychic" outcome. As expected, we find children whose explanations and predictions are consistent with each model, and an interaction between prediction and explanation. Critically, we find unexpected outcomes only induce children to move from predictions consistent with the simpler model to those consistent with the more complex one, never the reverse.

In everyday life we often 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. These abilities have been called *theory of mind* (Premack and Woodruff, 1978); theory of mind has 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 sees Sally place her toy in a basket, then go out to play. In Sally's absence her toy is moved to a box (causing her belief about the toy's location to be false). The subject is then 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 in the third or fourth year of life, from below-chance to above-chance performance (see Wellman et al. (2001) for a review and meta-analysis, though see also Onishi and Baillargeon (2005)).

It has 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 (i.e. beliefs are always consistent with the world) to a *perspective theorist* (PT) position (i.e. beliefs are mediated by perspective, and can be false). Perhaps the most striking comparison between these two theories is the asymmetry in their explanatory resources: the PT theory can make false belief predictions and explanations while the CT theory cannot.

Another influential thread of research 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 research on false belief (cf. Leslie, 1994; Gopnik and Wellman, 1992).

Empirically, the false belief transition is slow: children do not immediately achieve false belief understanding when exposed to evidence *prima facie* incompatible with the CT position (Amsterlaw and Wellman, in press; Slaughter and 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? That is, why would one ever accept a theory with fewer explanatory resources, unless perhaps, there is a drawback to the greater flexibility of the alternative theory? Indeed, intuition suggests that greater explanatory ability must come at the cost of greater complexity and, by Occam's razor, it should thus require additional evidence to accept the richer theory.

However, it is difficult from an informal description to know how explanatory resources and complexity differ between these theories, and how these factors should interact with evidence. Gopnik et al. (2004) have suggested that intuitive theories may be represented as causal Bayesian networks (Pearl, 2000); we use this framework to specify two models[1] of false belief. By applying Bayesian methods we investigate the rational transition between these models, balancing explanatory

[1]Our formal analysis takes place at Marr's computational level of modeling (Marr, 1982), that is, we describe the competencies, but not the algorithms (or processes), of cognition.

resources against complexity, and illuminate the above revision puzzle. To probe these ideas experimentally we investigate children's predictions and explanations, in cases when these predictions succeed and when they fail: the false belief task with the standard outcome (surprising to CTs), and a novel "psychic" outcome (surprising to PTs). We present only the apparatus necessary for a first investigation, leaving important elaborations for future work.

Formal Models

In the standard false belief task, described earlier, the story begins with Sally putting her toy in the basket. 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 (i.e. whether the door of the basket 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 goals and desires (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.)

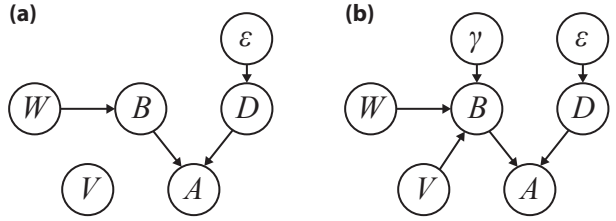

Figure 1: The dependency graphs of our Bayesian Network Models: (a) CT model, (b) PT model. Variables abbreviated by their first letter (see Table 1).

To specify the relationships between these variables we fix their joint distribution by giving a causal Bayesian network. The pattern of conditional dependencies, given by the directed graphs in Fig. 1, codifies the intuition that action is determined by beliefs and desires, and that belief is affected by the state of the world. In the PT model belief also depends on access[2].

The conditional dependencies are parameterized by the conditional probabilities given 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 $1 - \varepsilon$, 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 Access* is present, but without access Sally maintains her original belief, $Belief = 0$, with probability $1 - \gamma$. The 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 basket....

We assume asymmetric-beta priors on ε and γ. In the example simulations described below (Figures 2 and 3) 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). The relative magnitude of the two parameters determines whether it is more likely that Sally wants something other than her toy, or that she changes her belief – we have chosen the latter (because standard false belief tasks emphasize that Sally wants her toy). Otherwise, the qualitative results described below are quite insensitive to the values of these parameters.

Prediction

Having represented our models as probability 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. These models are *causal* theories: they also 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. 2(a)). There are two possible outcomes: Sally may look in the original location (the basket), or the new location (the box). We may predict the probability of each outcome by marginalizing the unobserved variables. Fig. 2(b) shows that the two models make opposite predictions. We see that the CT model "fails" the false belief test by predicting that Sally will look in the new (true) location, while the PT model "passes" by predicting the original location. The surprising outcome cases differ for the two models (looking in the original location for CT, looking in the new location for PT). Note that while the surprising outcome is not impossible in either model, it is far less likely in the CT model (as evident from Fig. 2(b)). That is, there is an explanatory asymmetry: *prima facie* equivalent unexpected outcomes weigh more heavily against the CT model than the PT model.

Theory Revision

Strong rationality requires an agent to balance the available intuitive theories against each other. How should a 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 its posterior probability given

[2]This is a simplification: we model how belief *content* depends on access, but it is likely that access mediates knowledge (vs. ignorance) even in the earlier theory.

Variable	Description	States
World (W)	Location of the toy.	0: Original location, 1: New location.
Access (V)	Could Sally see the toy moved?	0: No, 1: Yes.
Action (A)	Where Sally looks for her toy.	0: Original location, 1: New location.
Belief (B)	Where Sally thinks the toy is.	0: Original location, 1: New location.
Desire (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

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

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

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

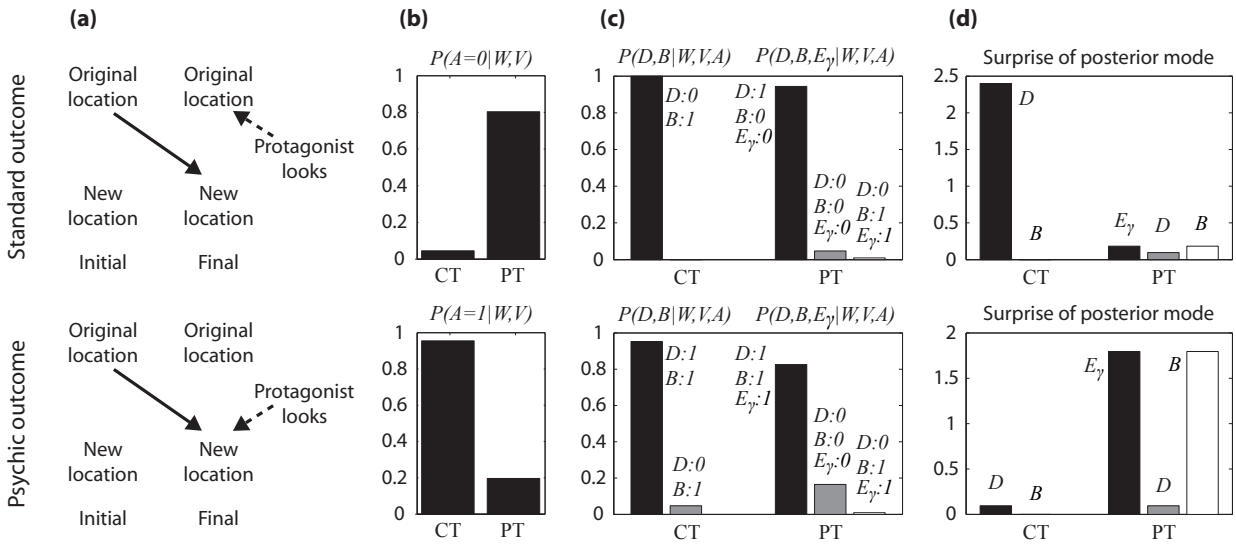

Figure 2: Comparing the Models: (a) Example situations, in both cases W=1, V=0, for the Standard outcome A=0, for the Psychic outcome A=1. (b) The predicted probability of each outcome. (c) Posterior probability of configurations of hidden variables, after observing the outcome. This indicates degree of belief in the corresponding complete explanation. (d) Surprise value of each variable in the modal configuration (computed as surprisal with respect to the predictive posterior).

previous experience. We may then write down a belief weight comparing belief in the PT model to the CT model:

$$W_{\mathrm{PT/CT}} = -\log(P(\mathrm{PT}|X)/P(\mathrm{CT}|X)), \qquad (1)$$

where X represents experience in previous false belief settings. When $W_{\mathrm{PT/CT}}$ 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 $W_{\mathrm{PT/CT}}$ to be strongly positive the agent behaves as a PT. In Fig. 3 we plot $W_{\mathrm{PT/CT}}$ 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 (e.g. 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. Since this rational revision depends on the particular character and statistics of experience, 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 (Jefferys and Berger, 1992): the PT model has additional complexity (the free parameter γ), which is penalized via the posterior probability. However, the data themselves are 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

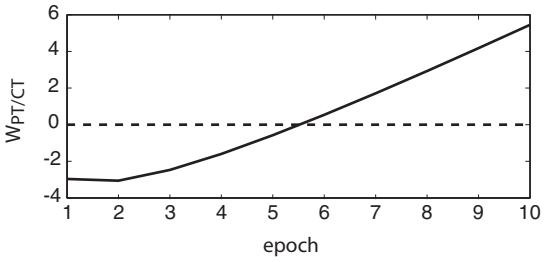

Figure 3: The log-posterior odds ratio over data epochs, showing the false belief transition from CT to PT. (Parameters integrated numerically by grid approximation.)

becomes favored.

When $W_{\mathrm{PT/CT}}$ is close to threshold, inferences will be mixed and far more sensitive to evidence[3]. In particular, any effect of the explanatory asymmetry noted above should be particularly prevalent in this period. When far from threshold, predictions in the situation of Fig. 2(a) will be largely consistent, but close to threshold they may become mixed, and such variability will be greater below threshold than above.

Explanation

As the above discussion emphasizes, an important function of intuitive theories is to explain observations by hypothesizing states of unobserved variables. To model this explanatory competence we first recast our models into a deterministic *explicit-noise form*, by introducing additional 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*.

Our PT model becomes deterministic if we introduce an *External Information* (or 'alternate access') 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, and has been omitted for clarity.) Explanatory inference is now dictated by the posterior distribution, conditioned on observations: the degree of belief in each complete explanation is given by its posterior probability (e.g. Fig. 2(c)). (See Halpern and Pearl (2001) for a related approach.) Once a complete explanation is chosen, a partial account of the explanans can be given by appealing to the *principle of surprise: a good explanans will address the ways in which an explanation is surprising*[4].

We can now characterize the explanatory asymmetry between the two models in more detail. Comparing the dependency structure of the two models, we see that access, alternate access (including external information), and belief (independent from the world) are causally relevant variables only for the PT model. We thus expect PT-theorists to appeal to belief and access more than CT-theorists, while CT-theorists will appeal primarily to desire. Indeed, to explain a surprising experience the CT model can only infer an alternate desire – Sally went back to the basket because she wanted the basket (not the toy). The PT model has additional explanatory resources in the sense that it can also appeal to access, particularly alternate forms of access, to explain unexpected outcomes – Sally went to the box because someone told her that her toy was there. Fig. 2(d) shows that these alternate desire and alternate access variables are indeed the surprising aspects of the most relevant complete explanations. From the principle of surprise we may then predict an interaction between theory and explanation type: in the surprising outcome cases CT-theorists will appeal to alternate desires, while PT-theorists will appeal to alternate access.

Children's Predictions and Explanations

If children are using intuitive theories of mind, as described here, then several model predictions should hold. First, there will be an evidence-driven false belief transition, and there will be a group of children near the threshold of this transition who exhibit decreased coherence in prediction and explanation. Because of the explanatory asymmetry between models, we predict that surprising outcomes will have greater weight before the transition than after. In particular, we expect some children who begin with CT predictions to switch to PT predictions when they encounter surprising evidence, but few children to switch from PT predictions to CT predictions when they encounter surprising evidence, and few children in any condition to switch when given consistent evidence.

Among children who are relatively far from threshold, and thus provide consistent predictions, there should be an interaction between prediction type and explanations: PT-predictors should generate more belief and access responses, while CT-predictors should generate more desire responses. Accordingly, we investigated the predictions and explanations that children generated when presented with the two possible outcomes to the standard false belief story. This is an extension to the prediction-explanation paradigm used by several authors to investigate false belief (e.g. Bartsch and Wellman, 1989).

Participants Forty-nine children (R=3;0-4;11, 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 were instructed not to interact with the child during the study.

Materials and Procedure Three picture books were created. The first book presented a guessing game unrelated to the false-belief task and was used to familiarize the child with the experimenter and with generating guesses. The other two books, Standard and Psychic, followed the standard Sally-Anne style narrative, with car-

[3]One expects the details of a process-level account to be especially critical near the threshold.

[4]Surprise may be formalized by the information-theoretic surprisal of a value with respect to some reference distribution, such as the predictive posterior.

toon pictures depicting the events throughout the book. These stories are equivalent to the situations of Fig. 2(a).

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. As a memory check 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?"

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. The order of the two test books was counter-balanced.

Results and Discussion Four children failed the memory test, and were excluded from further analysis. Based on the childrens responses to the prediction question in each book, we found three groups of children: 15 children who predicted the new location in both books (CT-predictors), 20 children who predicted the original location in both books (PT-predictors), and 10 children who changed their prediction based on the surprising evidence (Mixed-predictors). Critically, and as predicted by the explanatory asymmetry between the models, children only switched predictions after surprising evidence, and no children moved from the PT-consistent prediction to the CT-consistent prediction. Thus, of the children who initially made a CT-prediction, significantly more of those who received surprising evidence (for them the standard outcome) switched than of those who received confirming evidence ($p < 0.01$, $\chi^2 = 7.84$), and of the children who received surprising evidence significantly more who initially made CT-predictions switched than of those who initially made PT-predictions ($p < 0.01$, $\chi^2 = 8.60$). This order effect cannot be explained as a simple response to prediction failure, because the PT-predictors showed no increased tendency to switch when they received surprising evidence (for them the Psychic condition). One might worry that only the younger children were flustered enough by a wrong prediction to begin random guessing. This would imply that older children should be less likely to switch predictions than younger children. In fact, comparing the ages of the CT-predictors (M=3;7) to those of the Mixed-predictors (M=3;10) revealed that the Mixed-predictors were significantly older than the CT-predictors ($p < 0.05$, $t = 2.06$). This suggests that this pattern of mixed predictions is indicative of children who are close to the false belief transition.

Explanations were coded for the mention and value, if any, of each variable as in Table 1. (For the remainder, we have combined External Information with Access for clarity.) 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 Access=1. Another (CT-predictor) child explained the Standard outcome by "well, that's where she wants to look," which was coded Desire=0. Responses were scored by two coders, one who was blind to the group type for each child and to the formal model; inconsistencies were resolved by discussion. There were no significant differences between groups in references to observed variables (Initial World, Final World, Action).

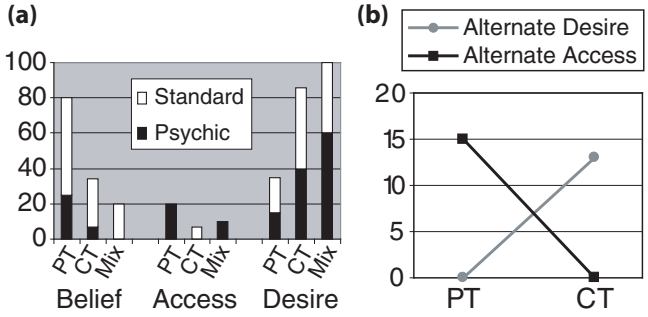

Figure 4: Summary of explanation data: (a) Portion of responses for each prediction group and condition referencing Belief, Access, and Desire. (b) Portion of responses asserting alternate values of Desire and Access, by prediction group.

As predicted, more CT-predictors than PT-predictors gave Desire explanations ($p < 0.025$, $\chi^2 = 5.60$, across conditions), and more PT-predictors than CT-predictors gave Belief or Access explanations ($p < 0.025$, $\chi^2 = 6.60$, across conditions). The Mixed-predictors gave explanations that were quite similar to the CT-predictors: their references to both Desire and Belief+Access were significantly different than the PT-predictors (both $p < 0.01$ by χ^2), but not the CT-predictors. Children's explanations are summarized in Fig. 4.

There is no way to know the chance level for explanations. However, consider that for an explanation to be coded as referring to alternate access or alternate desires, children had to spontaneously invent and report details outside the story (e.g. "he heard his sister move it", or "she wanted to move the basket"). It is thus suggestive that three of twenty PT-predictors gave alternate access explanations, and two of fifteen CT-predictors gave alternate desire explanations, in the respective surprising outcome conditions. Further, as predicted by our formal analysis, there is a significant interaction between prediction group (CT vs. PT) and type of the alternate explanations that were offered (desire vs. access) ($p < 0.05$ by 2×2 mixed ANOVA, $F(1, 136) = 5.04$). The interaction remains significant when restricting to the surprising

outcome conditions ($p < 0.05$ by 2×2 mixed ANOVA, $F(1, 66) = 5.31$). Interestingly, the Mixed-predictors offer significantly more alternate explanations (access or desire) than the non-mixed (PT and CT) predictors ($p < 0.025$, $\chi^2 = 5.02$).

Conclusion

The history of developmental psychology has been filled with 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 false belief reasoning 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.

Two of the primary advantages of formal models have been illustrated here. First, added precision can illuminate theoretical problems that resist simple solution. Indeed, our computational account sheds some light on the puzzle of rational theory revision by bringing the tools of Bayesian analysis to bear on the tradeoff between explanatory resources and complexity. Second, a model can suggest novel experimental avenues. Consideration of the formal structure of our models, especially the explanatory asymmetry between them, suggested designing an outcome condition which would be surprising to children who passed the false belief test – the novel Psychic outcome condition of our experiment. This condition then provided the crucial contrast needed to understand the interaction between theory and explanation, and to detect the outcome-order effect in predictions. These in turn suggest further experimental and theoretical avenues, such as a training study to test our suggestion that certain mixed predictions are a signature of children very near to the false belief transition.

The present account of the false belief transition 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 distributions 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), Rebecca Saxe, Tania Lombrozo, and Tamar Kushnir. This research was supported by the James S. 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*.

Jefferys, 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.

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.