Interactions between causal models, theories, and social cognitive development

David M. Sobel, David W. Buchanan, Jesse Butterfield, Odest Chadwicke Jenkins

A R T I C L E   I N F O

Keywords:
Causal models
Pretense
Learning from testimony
Markov random fields
Causal generalization
Rational Model of Categorization

A B S T R A C T

We propose a model of social cognitive development based not on a single modeling framework or the hypothesis that a single model accounts for children's developing social cognition. Rather, we advocate a Causal Model approach (cf. Waldmann, 1996), in which models of social cognitive development take the same position as theories of social cognitive development, in that they generate novel empirical hypotheses. We describe this approach and present three examples across various aspects of social cognitive development. Our first example focuses on children's understanding of pretense and involves only considering assumptions made by a computational framework. The second example focuses on children's learning from "testimony". It uses a modeling framework based on Markov random fields as a computational description of a set of empirical phenomena, and then tests a prediction of that description. The third example considers infants' generalization of action learned from imitation. Here, we use a modified version of the Rational Model of Categorization to explain children's inferences. Taken together, these examples suggest that research in social cognitive development can be assisted by considering how computational modeling can lead researchers towards testing novel hypotheses.

© 2010 Published by Elsevier Ltd

1. Introduction

By the time this paper reaches press, the home department of the first two authors will have transformed into a new department of Cognitive, Linguistic, and Psychological Sciences. A goal of the new department is to emphasize the multidisciplinary study of mind. This is wonderful in the abstract, but the challenge faced by members of this new department (and perhaps by all researchers interested in at least two of these fields) is that subfields in psychology, cognitive science, and computational modeling have gotten so specialized that it is not always the case that researchers approach questions in a way that truly touches on multiple disciplines, let alone contributes deeply to more than one area.

This appears to be the case when one considers computational modeling approaches to cognitive development in general issues in cognitive development, and the development of social cognition in particular. While many developmentalists use sophisticated statistical models to analyze data, there are only a handful of researchers using computational models to bolster developmental theories. We are agnostic as to why this appears to be the case. It might be due to the complexity of the models that are presented. It might be due to the occasional disconnect between the mathematical algorithms used to explain behavior and the empirical predictions made by those frameworks. It might also be more sociological: the history of studying cognitive development has witnessed several computational frameworks (e.g., symbolic reasoning, neural networks, dynamic systems, and most recently, causal graphical models and Bayesian inference) emerge as leading contenders for explaining behavior, each with seemingly as many supporters as detractors. For this reason, many developmental psychologists might discount the benefits of new computational frameworks until clear consensus in the field is reached.

Regardless of the reason that only a handful of developmental psychologists use computational models, our hypothesis is that much good can come from thinking through the interaction between these fields. The goal of this paper is to offer several examples of what we will call the Causal Model approach to modeling aspects of social cognitive development. Our thesis is simple: a modeling framework makes assumptions and predictions about human behavior that are testable. Investigating computational models to augment theories of development allows researchers to consider new predictions based on the nature of those models, the relation between the modeling framework and the developmental process, and the behavior predicted by the model. We hope that the examples we present below illustrate a way of thinking about building computational models of experiments and experimental paradigms with the goal of not only explaining existing data, but also generating novel, testable hypotheses, which can then be instantiated in new developmental experiments.

Two important caveats: first, this is not a novel idea in Cognitive Science. We borrow the name “Causal Models” from Waldmann.

* Corresponding author. Tel.: +1 401 863 3038; fax: +1 401 863 2255. E-mail address: David_Sobel_1@brown.edu (D.M. Sobel).

0893-6080/$ – see front matter © 2010 Published by Elsevier Ltd
doi:10.1016/j.neunet.2010.06.004
(1996), who suggested a similar account of incorporating logical and causal models into tasks involving adult judgment and reasoning.\(^1\) We posit that this account has deeper origins in philosophical investigations of logic and the lessons learned from the history and philosophy of science (e.g. Glymour, 1992). Second, we want to emphasize that it is doubtful that a single modeling framework constitutes the “best” framework for the purposes of modeling social cognitive development. In all of the examples that follow, it is certainly plausible that another modeling approach could explain the same data. What we are interested in is describing modeling projects that can interact with empirical research and lead to novel, testable hypotheses. In doing so, these projects support behavioral experiments that would not be conceptualized without the modeling framework.

What follows are examples of this approach. For each example, we will describe a phenomenon in psychological research, and then present a model of that phenomenon. Finally, we will discuss a novel empirical project (or empirical projects that are underway) that resulted from considering assumptions made by that framework serving as a computational-level description (cf. Marr, 1982) of children’s behavior. Thus, the remaining sections of this paper are organized such that any can be skipped if one is not interested in the particular phenomenon described.

### 2. Example 1: Children’s understanding of pretense

**A Phenomenon.** For many years, researchers in social cognitive development have considered what children know about pretending. Pretense emerges quite early in development (e.g. Piaget, 1962) and children understand the social nature of pretense before their third birthday (e.g. Harris & Kavanough, 1993). Critically, understanding that another person is pretending requires similar representational abilities as understanding that another person has a false belief—in both cases, the other person’s actions are contrary to what is expected given the actual state of the world (Leslie, 1987; Lillard, 1993a, 1993b). It is interesting that pretending emerges so early while children do not generally pass standard measures of false belief until approximately the fourth birthday (e.g. Wellman, Cross, & Watson, 2001).

There are two broad accounts of this discrepancy. The first is that young children recognize the representational aspects of pretending before they recognize the analogous representational aspects of belief (e.g. Perner, Baker, & Hutton, 1994). Consistent with this approach, several studies suggest that there are direct relations between children’s pretense and their success on other tasks involving mental representation, like standard measures of false belief. For example, the degree to which children engage in pretense (e.g. Astington & Jenkins, 1995; Lalonde & Chandler, 1995; Lillard, 2002), and the nature of that pretense, such as whether children have stable imaginary companions (e.g. Taylor & Carlson, 1997), predicts children’s false belief success. Similarly, preschoolers understand that characters who are pretending are thinking about their pretense (e.g. Br Dell & Woolley, 1998; Custer, 1996) and are trying to act like their pretense (e.g. Joseph, 1998; Rakocy, Tomasello, & Striano, 2004). A similar version of this account is offered by recent investigations on “implicit” theory of mind. Eye gaze studies on 3-year-olds’ performance on standard false belief tasks have suggested some early competence prior to children’s ability to make explicitly correct responses (e.g. Clement & Perner, 1994). More recently, Onishi and Baillargeon (2005) have demonstrated that 15-month-olds’ patterns of eye gaze in a violation of expectation paradigm reveal an understanding of others’ false beliefs. This view is consistent with Leslie (1987, 1988), who argued that the emergence of pretend play indicated that children possessed the metarepresentational abilities necessary to succeed on false belief measures. Indeed, using a similar violation of the expectation paradigm, Onishi, Baillargeon, and Leslie (2007) demonstrated that 15-month-olds understand the representational quality of others’ pretense. On this account, very young children have this representational capacity and fail standard false belief tasks because they lack inhibitory control or other executive function capacities necessary to give a proper response; these cognitive capacities develop around the fourth birthday.

The alternative account regarding the disparity between children’s pretense and success on false belief tasks suggests a disconnect between this early competence in pretense and children’s later representational abilities. Lillard (1993a, 1993b) argued that young children understand pretending as “behaving as if”. They base their judgments about pretending on another person’s actions rather than his/her mental states (see also Harris, 1991; Harris, Lillard, & Perner, 1994; Lillard, 2001; Perner, 1991). Support for this view comes from Lillard’s (1993a) “Moe the troll” task: 4-year-olds tend to judge that someone who is acting like an entity (e.g. Moe the troll, who is hopping like a kangaroo), but who does not know about the entity (e.g., because kangaroos do not exist in the land of the trolls, and Moe has never heard about nor seen one) is pretending to be that entity. These same children typically succeed on standard measures of false belief.

There have been various investigations contrasting a character’s actions with their mental states. All suggest that 4-year-olds have difficulty recognizing that characters are pretending when they act like the pretense entity, but are not engaging in a mental process that is a necessary condition for pretense (e.g. Ganea, Lillard, & Turkheimer, 2004; Lillard, 1998; Sobel, 2004, 2007). Studies have also suggested that 4-year-olds do not understand pretending as a mental state or as involving mental activity (e.g. Lillard, 1996; Sobel & Lillard, 2002). This line of investigations has suggested that an understanding of pretense as a mental state, related to and dependent on other mental states, develops between ages 7 and 8 (e.g. Lillard, 2001; Richert & Lillard, 2002).

A Model. Is it possible to reconcile these approaches? Some have argued that poor performance on the Moe task and other tasks like it are the result of children’s poor inhibitory control abilities (e.g. Frye, 2000). Sobel (2004, 2007) suggested that children’s developing success on these measures is not necessarily the result of developing executive function capacities. Further, this account does not explain why children begin to succeed on the Moe task around age 7, long after they succeed on standard measures of inhibitory control that have been thought to correlate with theory of mind measures (e.g. Carlson & Moses, 2001).

An alternate means of reconciling these approaches come from considering the representation of children’s knowledge of the causal relations among pretending, behavior, and other mental states. Consider a simple model of the Moe task, shown in Fig. 1. The figure is not meant to show a causal graphical model (or represent any specific modeling framework), rather to illustrate what the causal relations among mental states and behavior regarding pretense might be. The model shows three variables: (1) the intention to pretend to be an entity, (2) knowledge of that pretense, and (3) the action of the potential pretendor. The causal relation between the intention to pretend and the pretendor’s action might be construed as generative. An actor with the intention to pretend will act in a manner consistent with that.

---

\(^1\) Sloman (2005) also uses this nomenclature. He was mostly concerned with applying a particular modeling framework—causal graphical models—to measures in cognitive and social psychology. This is not our intention. As mentioned in the text, we do not want to endorse the use of any one particular modeling framework, but rather to take seriously the idea that any modeling framework can be used to generate testable empirical hypotheses. As such, we would argue that the phrase is simply polysemous.
entity (unless particular circumstances prevent this). The relation between knowledge of the pretense entity and the intention to pretend is not generative but rather an enabling condition; knowing about a kangaroo allows but does not compel one to pretend to be a kangaroo. Thus, this causal model relies on children recognizing both an enabling condition relation and generative causality. If children understand generative causality, but not enabling conditions, then the Moe task might underestimate children’s knowledge.

What do children know about enabling conditions? Toddlers remember sequences of events better if those events build on each other—if an early event enables a later event to be possible, rather than if the events occur in an arbitrary order (e.g., Barr & Hayne, 1996; Bauer, 1992; Mandler & McDonough, 1995; Wenner & Bauer, 1999). But better memory for sequences of events does not show that children specifically understand why certain sequences are easier to remember than others. Shapiro and Hudson (2004) found that preschoolers’ ability to plan and enact events was limited when those events involved many enabling condition relations as opposed to arbitrary relations. More generally, Siegler (1976) found that 5-year-olds have difficulty making inferences about necessity and sufficiency among causal events. Because an enabling condition is necessary for a generative relation to function, but insufficient to produce the effect by itself, these data suggest that young children might have difficulty reasoning about this kind of causal relation.

Hopkins and Sobel (2007) presented 4- and 6-year-olds with a novel causal environment in which children had to reason about a novel enabling condition relation. They showed children a machine that lit up and played music when certain objects were placed on it (the machine was actually controlled by a switch hidden from the child, which would cause it to always or never activate when an object was placed on it; the experimenter simply flipped the switch unbeknownst to the child at the appropriate places during the procedure). Children were shown a set of wooden blocks, each with a hole drilled into it, covered by a dowel, which when removed revealed a piece of white or yellow plastic inside, labeled as the “insides” of the objects. Children were told that these blocks came from different categories. The critical contrast was that “Tibs” never made the machine go, regardless of their insides, while “Tomas” made the machine go, but only if their insides were yellow (as opposed to white). This information was demonstrated to the child, with the aid of a second machine (which we will refer to here as the “charger”, although it was not given that label in the experiment). This machine changed the insides of the objects from white to yellow, via a false back and the appropriate sleight of hand on the part of the experimenter.

In this demonstration, children learned that the insides of the blocks acted as an enabling condition for one type of object. Having a yellow inside alone was insufficient to cause activation (i.e., children observed that tibs with yellow insides did not activate the first machine). Rather, if an object was a toma, then having a yellow inside enabled it to activate the machine. To test whether children understood this information, they gave them a test trial in which children saw two new objects with white insides. One was labeled a toma, the other a tib, and children were asked to activate the first machine using the two objects and the charger.

Six-year-olds mostly took the object that was labeled a toma, placed it in the charger to change its inside, and then placed it on the first machine. Four-year-olds in contrast, often did not do this initially, and were equally likely to place the toma or the tib in the charger first. This suggests that in this novel causal environment, 4-year-olds struggled to recognize the enabling condition relation from just the information they were shown. A Prediction. Perhaps 4-year-olds have difficulties with the Moe task result from a similar inability to recognize enabling conditions. If such an understanding was made clearer to young children, they might now demonstrate success on this task. In particular, there are contexts in which preschoolers seem to know about enabling relations. One is the relation between batteries and electronics: children often explain artifact motion by appealing to those objects’ batteries (Gelman & Gottfried, 1996), but they also generate and hear the word in everyday conversation.

Hopkins and Sobel (2007) also examined whether providing 4-year-olds with this contextual information might assist their reasoning in this novel causal environment. They introduced two new groups of 4-year-olds to a similar causal environment in which there were two kinds of objects—one that never made a machine activate regardless of its insides (akin to tibs) and one that only made that machine activate if its insides were in one state as opposed to another (akin to tomas). The manipulation here was that for one group of children, the insides were labeled “batteries”, while in the other condition they received the more

---

2 Sobel (2009) presented a brief CHILDES analysis (MacWhinney, 2000) on children’s use of this word. He examined transcripts of five children between the ages of 2:6 and 5:0. All children generated the word battery in their spontaneous speech in relation to making toys or other electronic devices work. All of the children also heard at least one of their parents talk about batteries, also in this context. As an example:

Abe (at 3:4, taken from Kuczaj, 1977):

CHI: I have another battery we could Daddy # I bought two batteries home.

FAT: Abe # I don’t think it’s the battery it’s probably the light bulb.

This example suggests that when the child observes a familiar causal relation fail to be effective, he appeals to the battery as a possible enabling mechanism that is not in place. The parental feedback gives the child confirmation that batteries could be a causal factor, but probably is not in this case.
generic label “insides”. They found a significant difference in 4-year-olds’ responses: when the inside of the objects were labeled “batteries”, these children responded like the older children in the previous study—they had little trouble recognizing the enabling relation. When the insides of the objects did not receive this label, 4-year-olds responded no differently from chance.

Sobel (2009) then showed that a similar manipulation could be done with the Moe task. He presented 4-year-olds with a troll doll (shown in Fig. 2). Children were told that the troll doll was “from the land of the trolls” and that in the land of the trolls there was one kind of animal (e.g., rabbits), but not another (e.g., kangaroos). In one condition, the animals were represented as pictures placed on the table, which the troll doll stood next to (i.e., he stood next to the picture of the rabbit, because he knew about them). In the other condition, the animals were represented by the same pictures, adhered to AA batteries, which were placed in the troll’s backpack. The troll was then made to act in a manner consistent with both of the animals (e.g., he hopped up and down). This was pointed out to the child, and the child was told that when he hopped up and down he looked like both animals. After a series of control questions to ensure children understood this information, they were asked whether the troll doll was pretending to be the animal that he did not know about (i.e., a kangaroo).

When children were shown the pictures placed on the table, children’s responses were not different from previous results by Lillard (1993a, 1993b, 2001)—they responded “no” (the correct response) ~35% of the time. In contrast, when the pictures were attached to batteries, children responded correctly ~70% of the time, significantly greater and significant more likely than chance. Sobel (2009) went on to show that this increase in performance was specific to the battery manipulation, and not due to other characteristics of the manipulation, such as potentially less inhibitory demands placed on the child in the battery condition.

Summary. Very young children might have some understanding of mental representation in pretense (Leslie, 1987, 1988; Onishi et al., 2007). However, in order to respond correctly on procedures like the Moe task, children must integrate that explicit knowledge with the other mental states involved in pretending (e.g., Lillard, 2001; Richert & Lillard, 2002). A way of resolving these findings might be that what’s developing during the preschool years is an understanding of the explicit causal structure (and the nature of those causal relations) among mental states like pretense and knowledge. Young children engage in pretense and thus understand that the intention to pretend involves other mental states, but the explicit understanding of how those mental states are related seems to develop later on.

3. Example 2: Learning from “Testimony”

A Phenomenon. A great deal of physical knowledge in the environment is learnable from observation and interaction with the world. Unsupported objects fall. Physical dimensions determine containment. Solid objects do not pass through one another and move together with cohesion. But mental states are different. We do not ever see a belief, desire, or pretense. We only observe behavior consistent with holding a belief, wanting an object, or pretending to be something. Children often learn about the meaning of words like “belief”, “want”, “think”, or “pretend” from listening to others. Of course, mental states are not the only domain where children have to rely on other people as sources of knowledge. Children rarely (if ever) directly observe true instances of certain biological entities (like germs) or magical creatures (like Santa Claus) yet often appeal to them in explanations of events. Any domain that involves conventional knowledge is not directly learned from observation and interaction with the world—such domains are learned from other people (e.g. Harris & Koenig, 2006).

But the arena in which the necessity of others’ input seems to be most important is learning the meaning of words. There is a conventional, but arbitrary assignment of lexical items to concepts (e.g. Bloom, 2000; de Saussure, 1966; Koenig & Harris, 2008). In order for children to learn the mapping for their specific language, they must come to recognize that others are reliable sources of knowledge and discriminate between the knowledge they acquire from a reliable and unreliable source. By the age of 3, children use a speaker’s reliability to make inferences about the extension of novel labels to novel objects (e.g. Clément, Koenig, & Harris, 2004; Koenig, Clément, & Harris, 2004; Koenig, Harris, & Harris, 2005; Pasquini, Corriuoy, Koenig, & Harris, 2007). These results suggest that children develop ontological commitments about events they cannot directly observe based on the “patterns of testimony” that they receive (cf. Harris, Pasquini, Duke, Asscher, & Pons, 2006).

One of the best examples of this experimental paradigm comes from Koenig and Harris (2005). They introduced preschoolers to two confederates (whom we will refer to as R and U for reliable and unreliable, respectively). An experimenter showed the child and confederates a set of familiar objects (i.e., objects that in pretesting, preschoolers did not err in labeling), and each confederate was asked to label the object. Confdrate R generated the label that children would have generated; confdrate U generated a label that children knew referred to a different object. After these familiarization trials, the experimenter showed the child and confederates novel objects (i.e., objects that in pretesting, preschoolers did not label consistently). Children were asked which of the two confederates they believed knew the label for this object. After children responded, the experimenter elicited a label from each confederate, and each person generated a novel label (e.g., one called it a dax, the other called it a wug). Children were asked to endorse one of the two labels. Koenig and Harris (2005) demonstrated that 3- and 4-year-olds asked and endorsed confiderate R significantly more often than confiderate U (and more often than chance). Children also inferred that an accurate labeler knows the intended function of a novel artifact than an inaccurate or ignorant labeler (see also Birch, Vauthier, & Bloom, 2008).

A Model. Butterfield, Jenkins, Sobel, and Schwertfeger (2009) proposed a model in which interactions between variables in a pairwise Markov random field (MRF) could be used to explain this phenomenon. A pairwise MRF is an undirected graphical model in which nodes represent either observed or hidden variables with interactions between them. An example, which we will use to explain the Koenig and Harris findings, is shown in Fig. 3.

Agents are represented by two nodes. The first (P1, n) represents each agent’s perception. The second (B1, n) represents
Fig. 3. Markov random field representation of the procedure used by Koenig and Harris (2005) to measure children’s inferences about speaker reliability. See the text for details.

Each agent’s belief and intentional states. These belief states govern the agent’s actions or the inferences that the agent makes based on the interaction with other agents and their own perception of the world. Each agent’s perception is fixed, based on the information available in the world. An agent infers information about other agents’ belief states based on their own belief state and communication with those agents, in the manner described below.

Each agent’s belief states are dependent on their perceptions, represented by an undirected edge that links these nodes. This connection is governed by a local evidence function \( \Phi(P_i, B_i) \), which dictates what that agent’s belief states should be given what he/she perceives. Put another way, the resulting distribution of \( B_i \), given \( P_i \), is what the agent’s belief states would be if they did not take into account communication from any other agent (i.e., if they had no theory of mind). Communication between two agents is represented by undirected edges between two agents’ belief nodes. This connection is governed by a compatibility function \( \psi(B_i, B_j) \), which represents the influence of agent \( j \)'s beliefs over agent \( i \)'s beliefs. Thus, at any given time, an agent’s actions are a function of the local evidence they observe and the communication they have with other agents. Over time, an agent learns a compatibility function with each other agent, based on the agent’s own beliefs and what communication they have had with those other agents.

This framework can be used to model the difference in inference between the reliable and unreliable speaker in Koenig and Harris (2005) in the following manner. Fig. 3 shows the child’s (C) belief state and perceptual state \( (B_c, P_c) \) as well as the belief and perceptual states of the two confederates \( (B_R, P_R) \) who will label \( R \) and \( U \) for “reliable” and “unreliable”, respectively. What is critical about the training is that it establishes differences in the compatibility functions between the child and the two confederates. By virtue of generating the same label as the child, confederate \( R \) would develop a different compatibility function than the compatibility function with confederate \( U \). The child learns that confederate \( R \) is likely to generate the same label that the child would generate herself, whereas confederate \( U \) would be likely to generate a different label from the child.

Thus, when children are shown the novel object at test, they step through the following inferential process (depicted in Fig. 4). First, the child is shown the novel object (indicated by the triangle in the leftmost panel). The child sees this object and her local evidence function indicates an equal probability that the object is a “dax” or a “wug”, as the child does not know its label (for simplicity, we will assume that these are the only two labels, but the model can be extended to include other possibilities).

Next (middle panel), the child is told each confederate’s label—\( R \) calls it a dax (d) and \( U \) calls it a wug (w). Because the child and \( R \) have consistently shared labels for familiar objects (objects where the local evidence function for the child indicated a high degree of probability on a particular label, which was the label generated by \( R \)), the compatibility function \( \psi(B_c = B_R) \) is comparatively high. Thus, \( R \)'s belief state should have a strong influence on the child’s belief states. The compatibility function for the child having the same label as \( U \), \( \psi(B_c = B_U) \), is relatively low (i.e., \( \psi(B_c = B_R) \gg \psi(B_c = B_U) \)). This allows the child to adjust her belief state so that more weight is given to the label “dax” and less weight is given to the label “wug”. This adjustment motivates the child to endorse the label “dax” (rightmost panel). Formally, this inference is shown in Eq. (1). The effect of the difference in compatibility functions between the child and the two confederates is that the child’s final distribution will be weighed more heavily by the reliable than the unreliable confederate’s belief state (in Eq. (1), the normalization constant \( 1/Z \) is present simply to ensure that the distribution sums to 1).

\[
Pr(B_i) = \frac{1}{Z} \phi(B_c, P_c) \sum_{i=R,U} \psi(B_c, B_i) \phi(B_i, P_i),
\]

A Prediction. This MRF framework qualitatively explains children’s use of speaker’s reliability in learning novel labels from others. An open question is how the model (and thus how children) define their compatibility functions. Butterfield et al. (2009) described two factors that defined the influence of compatibility on an agent’s belief states. First, higher levels of compatibility with a confederate were defined by a strong likelihood that the data a confederate generated was equivalent to what the agent’s local evidence suggests should be generated (i.e., their level of reliability). Second, the degree of confidence a confederate has in his own beliefs indicates more influence on the agent’s beliefs if the agent’s local evidence function does not specify a unique answer. This second process allows the MRF framework to explain data from (for example) Sabbagh and Baldwin (2001), who showed that preschoolers remembered novel labels for novel objects better if the same speaker generated that label with certainty than with uncertainty (Butterfield et al., 2009, also describes this simulation).

This model, as formulated for these experiments, assumed two processes: (1) all agents have the same baseline level of reliability and (2) when agents communicate with each other, that communication is always relevant to inferences at hand. Several studies, however, suggest that the first assumption might not apply to children, as they do not initially view all informants as equally reliable. Instead, children use knowledge about the confederates to infer reliability. For instance, children view an adult as more reliable than a child, even if both individuals are accurate (Jaswal & Neely, 2006). Children also view a familiar informant as more knowledgeable than an unfamiliar one (Corriveau & Harris, 2009), and trust the information provided by their mother based on their attachment relationship with her (Corriveau, Harris, et al., 2009). Finally, children trust an adult with no history of labeling familiar objects over an adult who is inaccurate when generating familiar labels (Corriveau, Meints, & Harris, 2009). What children know about the confederates as sources of knowledge affects how they interpret the data they generate. Obviously, in future simulations, we can update the model to account for these data.

But more importantly, the MRF framework can make predictions about the relevance of the information confederates tell to a child. When an agent (i.e., the child) is told a piece of information, she has a representation of her belief states about that information, which is based on the perceptual information she observed. That representation might be structured and coherent, in which the child believes something about the world (for example, that certain objects have particular labels, related to their category membership and kind). Alternatively, that representation might be an undifferentiated, incoherent set of facts, such that the child does

---

3 This computational framework was developed initially for robots acting cooperatively (Butterfield, Gerkey, & Jenkins, 2008), and thus these assumptions are justifiable for the problems it was initially intended to solve.
not know what events are connected by causal relations. In the former case, when confederates communicate irrelevant information to the child, she will recognize that irrelevance and will discount the reliability of the confederates regarding those data. In the latter case, because the child might not have a coherent theory, she might treat the more accurate (i.e., more positive) informant as being more accurate across domains of knowledge because the child does not know to discount this information. Qualitatively, this latter situation is similar to the child using the Gricean principle of informativeness (Grice, 1975); the child assumes that the information is relevant to the inference because why else would the experimenter be communicating it. That communicative intent is presumably relevant unless the child specifically knows it is not (based on her initial belief state, the perceptual information present, and the local evidence relating those percepts to the child’s beliefs).

A simple test of this prediction was performed by Sobel and Corriveau (2010), who introduced 4-year-olds to an experimenter and two confederates (who we will refer to as R and G for reasons that will become obvious below). The experimenter brought out a set of wooden blocks one at a time. One group of children was randomly assigned to be in the inside condition; these children were told that the objects were either made of red stuff or green stuff (represented by plastic beads inside the object, hidden behind a removable opaque “door”). For each object, the experimenter asked each confederate what she thought the object was made of. Children observed that confederate R always accurately knew when an object was made of red stuff, and said “I don’t know” for objects made of green stuff; confederate G always accurately knew when an object was made of green stuff, and said “I don’t know” for objects made of red stuff. Thus, both confederates were equally accurate, but appeared to have different expertise regarding what these objects were made of. The other group of children were in the outside condition; they were told that the objects either had red or green stickers on the back (the objects were positioned in such a way that only the experimenter could see their backs). Again, confederate R was always accurate regarding objects with red stickers on them and ignorant for objects with green stickers, while confederate G showed the reverse knowledge.

At test, children were shown a new object made of red/green stuff (in the inside condition) or with a red/green sticker on the back (in the outside condition). They were then told that the object had a special name, and each confederate was prompted to give a label. The confederates generated different novel labels (e.g., one called it a “dax”, the other a “wug”), and children were asked to endorse one of the novel labels generated by the two confederates.

In both conditions, the confederates displayed the same level of reliability. However, the relevance of that reliability differs across the conditions. Knowledge of what an object is made of clearly relates to that object category membership, and hence knowledge of its label (e.g. Gelman, 2003; Gottfried & Gelman, 2005; Sobel, Yoachim, Gopnik, Meltzoff, & Blumenthal, 2007). In contrast, whether an object has a red or green sticker on it is an arbitrary property, and only meaningful if it relates to some kind of conventional knowledge (i.e., stickers on the outside could indicate an object’s non-obvious insides, but that would require a prior agreement among communicative agents to be meaningful). As such, Sobel and Corriveau predicted that children would only use the appropriate expert’s knowledge in the inside condition. This was exactly what they found: 4-year-olds were included to endorse the red expert’s label for the object made of red stuff and the green expert’s label for the object made of green stuff, significantly more so than what would be expected by chance responding, and significantly more so than the expert’s labels in the outside condition.

A Second Prediction. In addition to predicting whether children will use the reliability information they observe judiciously, the MRF framework suggests an explanation for a set of divergent findings in the literature on speaker reliability. Some have argued that speakers’ history of reliability transfers to accuracy in many other domains—consistent with the hypothesis that children show a halo effect towards one of the speakers. Brosseau-Liard and Birch (in press) demonstrated that 5-year-olds (but not younger children) would state that a speaker with a history of accurately labeling objects would be more knowledgeable overall and would possess more prosocial attributes (e.g., the accurate speaker would be more likely to share).

The evidence for such “halo effects” is mixed. Fusaro, Corriveau, and Harris (2009) trained a group of 4-year-olds that one informant was stronger than the other and found that children believed the stronger individual was not only more likely to be able to lift a heavy object than the weaker individual, but also would be a more reliable source of knowledge about the labels of novel objects. In contrast, 4-year-olds trained that one informant was a more reliable labeler of familiar objects than another would endorse that individual’s labels for novel objects, but were at chance when asked which character could lift the same heavy object. A halo effect was observed in one direction—from strength to word learning, but not in the other.

The MRF framework can nicely explain this asymmetry. When preschoolers were trained that the confederates were accurate or inaccurate labelers of familiar objects, they did not use that information to determine who was stronger. Children’s more differentiated theory of category membership (i.e., kinds) allowed them to know that one’s ability to label accurately was unrelated to strength. This also explains why past reliability in object labeling should generalize to inferences about novel objects’ functions (e.g. Birch et al., 2008; Koeng & Harris, 2005); preschoolers often relate these two pieces of information, making inferences about one based on the other (e.g. Bloom, 1996; Gelman & Bloom, 2000; Kemler-Nelson et al., 1995; Kemler-Nelson, Frankenfield, Morris, & Blair, 2000). Children’s differentiated theory of kinds allows them to recognize the relevance of this information.

In contrast, children preferred to learn novel labels from a physically strong as opposed to a weak informant. It is likely that children treat strength as an undifferentiated concept.
Wenham (2005), for instance, suggested that elementary-school children often incorrectly conceptualize strength as just the ability to act with force (and not related to intrinsic properties of the object/individual). Thus, when preschoolers learn that one character is strong and the other is not, they do not know that this information is unrelated to word learning; they might assume a halo because it explains why they observed the information and trust the strong character when learning object labels.

Summary. The use of the MRF framework to describe aspects of theory of mind was initially conceptualized as a model of collaborative actions among robots, concerned with picking up information from another’s gaze (Butterfield et al., 2008). Our more recent work (Butterfield et al., 2009) suggests that this framework nicely explains (at least qualitatively) what information can be learned from interaction with others. One thing that this example suggests is that there are ways in which the psychological literature on testimony can clearly influence how these models are constructed, which is important for advancing how this modeling framework is used (particularly in more applied fields like robotics).

We recognize that in this example we have not offered as much of a mathematical description of MRFs and our models as one might like (and refer the reader to Butterfield et al. (2009) for that level of detail). Our point in doing so is to suggest that the model is acting much like a psychological theory in this case—offering a coherent account of why we should expect certain findings in the testimony literature, such as when children should invoke a halo effect and when they should not.

4. Example 3: Imitation and causal generalization

A Phenomenon. At the youngest ages, infants reproduce the behaviors of others (e.g. Meltzoff & Moore, 1977, 1983). But what are infants doing when they imitate action, particularly when they imitate action that brings about a causal relation? When and how do children learn causal knowledge from imitating actions? Sometimes during the second year of life, infants’ imitation is clearly based on understanding the intentional actions of others (e.g. Gergely, Bekkering, & Kiraly, 2002; Meltzoff, 1995) and does not involve “blindly” (cf. Want & Harris, 2002) reproducing the means of an action. Rather, infants develop an understanding that there is a relation between the action they observe and its efficacy (Elsner & Aschersleben, 2003).

However, most of this work only focuses on infants’ ability to link the actions they observe with the effects of an object in the context of that object and not whether they understood the actions to generate those effects in general. In a series of studies, Emily Bushnell and colleagues (e.g. Brugger, Lariviére, Mumme, & Bushnell, 2007; Yang, Bushnell, & Sobel, submitted for publication; Yang, Sidman, & Bushnell, 2010) have suggested that infants appreciate the causal nature of an action’s efficacy when they reproduce goal-directed behaviors modeled by others. Further, they argue that infants appreciate the action’s efficacy in general and not specific to a particular object.

To illustrate this, Yang et al. (submitted for publication) presented 15-month-olds with novel toys with individual handles that elicited pushing or pulling behaviors (shown in Fig. 5). In their first experiment, infants observed two toys during the demonstration phase of the procedure, which were identical except that they possessed different handles (one had a red button that depressed; the other had a yellow lever that could be pulled out). The experimenter acted on the first toy, which produced a novel effect (e.g., a rabbit would pop out of the pink can when its pull lever was manipulated). Children were then given the opportunity to imitate this action, and subsequent result. Then infants were shown a second toy that looked identical, but had a different handle, and the experimenter demonstrated that manipulating this handle had no effect on this new toy (e.g., pushing the red button on the same pink can would result in no effect). Again, infants were allowed to imitate this action, and lack of effect.

Only infants who imitated both actions were then moved on to the test trial (the majority of infants did so, and interestingly, they did not differ in the amount of time they spent imitating the effective and ineffective actions). In the test trial, children were shown a third toy, which looked identical to the demonstration toys, except that it possessed both handles. Infants were simply given this toy and allowed to play with it for 30 s. The majority of infants (69%) first touched the previously effective handle, and even more (75%) manipulated that handle first. Further, Yang et al. set this test toy to be ineffective, regardless of what handle was manipulated. This way, they could examine how persistent children would be in manipulating each handle. They found that infants manipulated the previously effective handle significantly longer than the previously ineffective handle.

Infants’ generalization abilities, however, do not appear to be terribly robust. Yang et al.’s second experiment demonstrated that 15-month-olds would not generalize the causal action when the test toy had the same handles, but was a different shape (see Fig. 5(a)). Instead of acting first on the effective handle, they responded at chance (touching and manipulating each handle first 50% of the time).

Their third experiment was inspired by the possibility that a particular computational model (which we will describe below) could help describe how children linked the data they observed during the demonstration phase of the experiment to the inference they were asked to make when shown the test toy with two handles. In this experiment, they replicated the findings of Experiment 2 in a new group of 15-month-olds (in what they called the same demonstration condition). Again, when shown that the handles differed on the same two demonstration toys, infants would not generalize the efficacy they observed to a test toy of a different shape with those same handles (touching and manipulating the effective handle 42% of the time in this replication). They also found that a new group of 15-month-olds would generalize the efficacy of the actions when different toys with different actions were used in the demonstration phase (shown in Fig. 5(b)). When infants were presented with different action handles on different objects (the switch demonstration condition), they were more likely to recognize that the action was not specific to the object itself, but could generalize the effects of those actions to new objects. In this case, they touched the effective handle first 67% of the time and manipulated that handle first 71% of the time, significantly more often than in the same demonstration condition, and more often than chance levels.

An interesting difference, though, between the results of the switch demonstration condition and the results of the first experiment (in which the three objects were all the same shape), was that when the overall amount of time spent playing with the effective handle was measured, children differed between these experiments. In Experiment 1, children overwhelmingly touched and manipulated the effective handle during the 30 s they were allowed to play with the test toy, significantly more often than they touched or manipulated the ineffective handle. This was not the case in the switch demonstration condition, where infants touched and manipulated both handles an equivalent amount of time during their free play. Yang et al. (submitted for publication) suggest that many infants in this condition touched and manipulated the effective handle first, observed that it was ineffective, and then switched to touching and manipulating the other handle. In contrast, many infants in their Experiment 1 persisted in touching and manipulating the effective handle, not accepting that it was ineffective.
A Model. We modeled these data using the Rational Model of Categorization (hereafter: RMC), proposed by Anderson (1990) and developed by Sanborn, Griffiths, and Navarro (2006). In this model, entities are assigned sequentially to categories, balancing two goals: (1) try to assign objects to categories that already have many entities in them and (2) try to assign objects in such a way that features are homogenous within categories.

We begin with the assumption that categorization drives infants to behave differently in the conditions described above. If infants categorize all three objects as members of the same category, they will be more likely to generalize the effectiveness of the handles to the test object and choose the effective over the ineffective handle. If they do not group the three objects as members of the same category, they will be less likely to do so and more likely to guess which handle to choose.

There are two parameters in the model: the first is $\alpha$, which captures the tendency for objects to be placed into the same category. The second is $\beta$, which captures how much the model prefers that categories have homogenous features. We will refer to the objects children see during training in Fig. 5(a) as $A_{1+}$ and $A_{2-}$ as they share the same shape ($A$), but differ in whether they possess handles 1 or 2 and whether they are effective (+) or ineffective (−). The objects in Fig. 5(b) are $A_{1+}$ and $B_{2-}$ (as they differ in shape, handles, and efficacy). $T_{12}$ refers to the test object in both conditions. Yang et al. counterbalanced whether the effective toy was first in the demonstration, so children observed combinations of either $A_{1-}$ and $A_{2+}$, or $A_{1+}$ and $A_{2-}$.

In the fully rational version of the RMC, the first object children observe (e.g., $A_{1+}$) is placed into the first category by default. The second object is placed either into the same category as the first object, or into a new category. This is more likely when more of the features match (i.e., when children are shown $A_{1+}$ and $A_{2-}$) than when they do not ($A_{1+}$ and $B_{2-}$), but is still possible in the latter condition. The strength of this preference depends on how much stronger $\beta$ is than $\alpha$: how much the model cares about feature homogeneity as opposed to category size. Mathematical details for this assignment are provided in the Appendix. The test object is treated in a similar way, depending on the features, $\alpha$ and $\beta$. In fully rational versions of the model, steps are also taken to ensure that order does not play a role: for instance, objects are randomly assigned in different orders. However, we will suggest below that order is relatively important.

In the cases where the first two objects are treated as members of the same category, how should the third object be treated? It depends on whether those objects share a common feature. Intuitively, when the first two objects share a common feature, the test object is facing a category that is more homogenous than when the first two objects are placed in the same category and do not share a common feature. Thus, the test object is more likely to be categorized as a member of the same group as the first two objects when those objects do not share a common feature than when they do.

To capture this intuition, we use a hierarchical learning framework (following Kemp, Perfors, & Tenenbaum, 2007). The model learns not just what categories the objects are in, but also the parameters with which it is working. If the model has the category formed by $A_{1+}$ and $B_{2-}$, it should learn that $\beta$ is strong: in this context, categories are more homogenous. If the model has learned the category formed by $A_{1+}$ and $B_{2-}$, it should learn that $\beta$ is weak—categories can include a wider variety of features.

We used a version of the RMC algorithm that was designed for this procedure. In this version, the model automatically assigned the first two objects to the same category, and then adjusted $\beta$ depending on the feature homogeneity of that category. We chose not to have the model learn $\alpha$: it did not help to capture the difference between conditions, only added unnecessary complexity to the model. We used a Monte Carlo simulation, generating 10,000 samples, and measured the proportion of samples on which the model assigned all three objects to the same category (see the Appendix for details). When the first two objects shared a common feature (i.e., shape), the model categorized all three objects as members of the same category 41% of the time. When the three objects were all shaped differently, it categorized these objects as members of the same category 52% of the time.

**Fig. 5.** a and b. Stimuli used by Yang et al. (submitted for publication) to examine toddlers’ causal generalization abilities from imitation. (a) An example stimulus set from the “same” condition of their Experiment 3, while (b) is an example stimulus set from the “switch” condition of that experiment. Children were more likely to manipulate the previously effective handle in the switch than in the same condition. These data are the basis of the model presented in the subsequent section.
These values are qualitatively similar to the “touch first” data from Yang et al. (submitted for publication, Experiment 3). In the same demonstration condition, children touched the effective handle first on the test object 42% of the time, in comparison to 67% of the time in the switch demonstration condition.

We contrasted this model’s performance with two alternate versions of the RMC model. The first was a direct application of the fully rational RMC, which was similar to the model above (i.e., it learned $\beta$) except that it did not assume that the first two objects belonged to the same category. Instead, it randomly reassigned objects based on the assignments of the other objects, using Gibbs sampling (details are provided in the Appendix). This model did not fit the data well: the three objects were categorized as members of the same category 27% of the time when the first two objects shared features, and 20% when they did not. Of course, under this model it was possible that the test object was co-categorized with only one of the test objects. This happened 23% of the time in the same demonstration condition and 34% in the switch demonstration condition. While this difference is in the correct direction, in order to fit the data qualitatively, we would have to assume that the children had a stronger tendency to choose the previously effective handle when the test object was co-categorized with only one of the training objects, than when all three objects were co-categorized, which seems incoherent.

The second alternate model was identical to the model above, but did not use a hierarchical learning framework. It automatically assigned the first two objects to the same category, and then chose whether to assign the test object to that category, but did so without adjusting the value of $\beta$. This model did not differentiate between cases in which the two demonstration objects shared or did not share a feature: it assigned the test object to the same category in both cases 55% of the time.

A prediction. The model assumes that the infant assigns the two demonstration objects as belonging to the same category. Is this a plausible assumption? It is certainly more plausible in the same demonstration condition than the switch demonstration condition, because the objects look alike, even though they do not share efficacy. However, infants might be making a deeper social inference: they might assume that the experimenter demonstrated the objects in a particular order for a reason.

A simple prediction arises from this: we should be able to change children’s generalizations by changing the manner the objects are presented. We are currently investigating a replication of the same demonstration condition in one of two situations. Half of the infants observe the two demonstration objects and then the test object sequentially (as in Yang et al.). The other infants initially observe all the three objects at once, but the training and test will proceed in the same manner. We would expect to see more generalization if all three objects were observable at the start of the procedure, as it might suggest that all objects come from the same category.

A second prediction. Recall that in Yang et al.’s (submitted for publication) experiments, the test toy was not efficacious, regardless of what handle was manipulated. However, Yang et al. observed a different pattern of manipulation between their first experiment (where all three objects were the same shape) and the switch demonstration condition of Experiment 3 (where all three objects were shaped differently). The model makes a prediction about these results.

Under our assumptions, persistence in manipulating the previously seen effective handle is indicative of a stronger belief that the test object is a member of the same category as the demonstration objects. We used the same model as above (e.g., it automatically assigned the first two objects to the same category, and learned $\beta$), but added a feature that represented the efficacy of the supposedly effective handle. We asked the model to estimate the probability that all three objects were in the same category, given the information that this additional feature did not match. We considered two cases—one in which all three objects shared the same shape (analogous to Yang et al.’s Experiment 1) and another in which all three objects were different shapes (analogous to the switch demonstration condition in Experiment 3). In both cases, children generalized to the effective handle—they touched and manipulated it first more often and more often than chance. However, the results of this simulation suggest that children’s consideration of the failure of the previously seen effective handle changes category membership differentially. When the three objects share the same shape, the drop in the probability that all three objects are in the same category is only moderate, from 80% to 62%. When all three objects are of different shapes, the drop is more drastic, from 51% to 26%.

This difference is measurable in the amount of time infants spend manipulating the previously seen effective handle during the 30 s they are given during the test trial. When all three objects were of the same shape, infants spent much more time manipulating this handle (and most of the time spent manipulating the handles overall). When all three objects were shaped differently, infants spent much less time manipulating this handle and were more likely and quicker to switch to manipulating the other handle. This analysis was suggested, and nicely explained, by considering the results of this model.

Summary. Using the rational Model of Categorization, we explain performance on a measure of causal generalization from imitation. When we assume that children are co-categorizing the first two objects (possibly a rational social inference) and adjusting their expectations of category homogeneity as they go, then we can explain why they seem to generalize in the switch condition, but not in the same condition. This model explains an additional aspect of the data: children appeared more resistant to additional evidence suggesting that the test object was not a member of the category when all three objects share common shape, than when all three objects differed in shape.

A final note: there is a similarity between the present results and other findings on infant categorization. For example, Quinn, Eimas, and Rosenkrantz (1993) found that children were more likely to dishabituate to a novel category when habituation was done with a diverse category than with a relatively narrow category; infants dishabituated to dogs when habituated with cats, but not vice versa. Quinn used a connectionist model to explain this difference, but there is no intrinsic contradiction between this approach and the present (more computational-level) approach we have presented here (Mareschal, Quinn, & French, 2002). We suspect a similar connectionist model could be constructed to account for the present data. An open question is whether it would make the same predictions, or completely different ones. This is clearly worth investigating.

5. Conclusions

The three examples we have generated consider lines of research integrating computational modeling with issues in social cognitive development. The first example illustrated that by just considering the architecture of a modeling framework (i.e., what is the nature of the causal relations among mental states), we could generate predictions about children’s understanding of aspects of their developing theory of mind. In this case, the “behaving-as-if” hypothesis of children’s understanding of pretense (e.g., Lillard, 1993a, 1993b, 2001) might be restated as children lacking an understanding of the exact causal structure among particular mental states, like the intention to pretend and the knowledge of the pretense. An open question in this area is whether other enabling condition relations among mental states develop in
conjunction with children’s success on ordinary versions of Lilhard’ Moe task. For example, understanding that attention to material, as opposed to simply the desire to learn material is necessary for learning appears to develop along a similar trajectory (Sobel, Li, & Corriveau, 2007). Does an abstract understanding of enabling conditions underlie this development? The second example is perhaps the most direct test of our Causal Models approach. Here, a modeling framework designed for robotics appears to explain a set of findings in social cognitive development—when we should expect to generalize information from a reliable individual or unreliable individual. Again, the predictions made by considering this model come from thinking through how the model is making inferences—in this case by comparing the agent’s own belief states with what they integrate from others’ reliability and the relevance of that information. The predictions here—such as when to expect a halo effect in generalization from reliability—are still being tested, but offer a good example of how models can act as theories in that they produce new empirical investigations.

The final example is not an example about social cognitive development but rather how social cognitive development does not act in isolation: learning from imitation is dependent on what information is being demonstrated to the infant. The model we presented is a modified version of a categorization model, tailored to developmental data. The first prediction, which allowed us to consider further empirical data that are being collected, came from testing the assumptions we made to fit the experimental procedure to the model. More importantly, the model also made a prediction about a way of considering how the existing data were analyzed—one that would not have been difficult to interpret without the model’s results as a guide.

Taken together, our hope has been to demonstrate that one of the most useful aspects of computational modeling is to help generate testable hypotheses. We also recognize a drawback in our approach, which we hope is a point for discussion. Our examples emanate from trying to predict (or predict retrospectively) aspects of human data. This might not be the same as explaining why human beings (whether adult or child) act the way they do. Many psychologists in general (and developmental psychologists in particular) are interested in explaining why we behave the way we do. The opposing argument is that in the cases we have generated (and we suspect many others in the literature), it is not always clear researchers would have discovered the behavior without the modeling components. A given model may or may not provide a satisfactory explanation of the behavior it models. We (following many others) have intended these models as computational-level descriptions of behavior, which may not yet provide a complete explanation. We hope they are a step toward eventually discovering and describing the algorithmic level, and how the brain instantiates those algorithms.

Acknowledgements

The research described in this manuscript was supported by NSF DLS-0518161 to DMS, an NSF Graduate Research Fellowship to JB, and ONR Award N000140710141 and AFOSR Award FA-9550-09-1-0206 to OCJ. We would also like to thank Noah Goodman and Jennifer Van Reet for helpful discussion about issues in this manuscript.

Appendix. Mathematical details for the RMC Used in Example 3

We compared infants’ performance to three different versions of the RMC. One used the fully rational Dirichlet process as described by Sanborn et al. (2006) and also incorporated hierarchical learning of $\beta$. The other two models were modified versions of this model. We will first describe the algorithms that each model used to sample the category assignments, and then specify the exact equations we used.

The fully rational model used Gibbs sampling. The objects were initialized to random clusters. At each step, we re-sampled $\beta$ taking the cluster assignments as given, then resampled each cluster assignment taking $\beta$ and the other cluster assignments as given. We sampled $\beta$ using an independence chain Metropolis–Hastings sampler: we proposed new values of $\beta$ from Exponential(1), then accepted them with a probability proportional to the likelihood ratios of the current and proposed values of $\beta$. One we had iterated the Metropolis–Hastings component at least ten times with at least one acceptance, we proceeded with the rest of the Gibbs sampling step. We repeated the Gibbs sampling until we had 10,000 samples, sampling the table assignments every 10th step.

The other two models were modified versions of this sampler, in which the first two objects were always assigned to the first cluster; their cluster assignments were not re-sampled. The only values sampled were $\beta$, which was only learned from the cluster assignments of the first two objects, and the cluster assignment of the test object. We repeated this 10,000 times and counted the proportion of times that the objects were all categorized together. One version was also without hierarchical learning—this was the same, except that we did not sample $\beta$; it was always set at 1.

To model the effect of additional feature data, we added a feature that represented the effectiveness of the previously effective handle.

We will now describe the equations used in the samplers. To do the Gibbs sampling, we first needed to know the probability that an object was a member of a cluster, given the parameters, features, and other cluster assignments. The RMC assumes that features are independent given the category. So, we can compare the likelihoods of each cluster assignment using the following equation:

$$P(C_i = c | C, F, \alpha, \beta) \propto P(C_i = c | \alpha)P(F_i = f | C_i = c, \beta)$$

where $F$ is a vector of features, and $C$ is a vector of cluster assignments, and $\alpha$ and $\beta$ are parameters. We set $\alpha = 1$. In this equation, and all of what follows, $C_i$ refers to all members of $C$ except $C_i$ while $C_{-i}$ refers to all $C_i$ such that $k < i$.

The categorization process was a Dirichlet process, which assigned a probability to each cluster as follows:

$$P(C_i = c | C_{-i}, \alpha) = \begin{cases} \frac{\#(j : C_j \in C_{-i}, C_j = c)}{\#(j : C_j \in C_{-i}) + \alpha}, & \text{if } c \text{ is an existing category} \\ \frac{\#(j : C_j \in C_{-i}) + \alpha}{\#(j : C_j \in C_{-i})}, & \text{if } c \text{ is a new category} \end{cases}$$

where $C_{-i}$ is the vector of category assignments of all previous objects, and $\#(j : C_j \in C_{-i})$ counts the number of previously categorized objects, while $\#(j : C_j \in C_{-i}, C_j = c)$ counts the number of previously categorized objects that were placed into category $c$.

The features were assumed to come from a multinomial with a Dirichlet prior, with $n_f$ possible states that were always equally likely:

$$P(F_i = f | C_i = c, \beta) = \frac{\#(j : C_j = c) + \beta}{\#(j : C_j = c) + n_f \beta}.$$

For the initial model, there was only one feature with $n_f = 3$ states, corresponding to the three shapes. In one version of the model, we added a binary feature ($n_f = 2$) corresponding to the efficacy of one of the handles.
We also learned $\beta$ by comparing the likelihood ratios of two candidate values. Each likelihood was proportional to

$$P(\beta = x | C, F, \alpha) \propto e^{-x} \prod_j \prod_i \left( \frac{\#(j : C_j^i \in c^{-i} F_j = f^i + x)}{\#(j : C_j^i \in c^{-i}) + n_j x} \right)$$

where $C_{-j}$ was the set of previous members of that category. So, the likelihood was proportional to the product of the probabilities of generating each feature sequentially using $\beta$, for each feature and each category.

References


