

## BAYESIAN SPECIAL SECTION

# Using domain-general principles to explain children's causal reasoning abilities

James L. McClelland<sup>1</sup> and Richard M. Thompson<sup>2</sup>

1. Department of Psychology, Carnegie Mellon University, Pittsburgh, USA

2. Department of Psychology, University of Exeter, UK

### Abstract

*A connectionist model of causal attribution is presented, emphasizing the use of domain-general principles of processing and learning previously employed in models of semantic cognition. The model categorizes objects dependent upon their observed 'causal properties' and is capable of making several types of inferences that 4-year-old children have been shown to be capable of. The model gives rise to approximate conformity to normative models of causal inference and gives approximate estimates of the probability that an object presented in an ambiguous situation actually possesses a particular causal power, based on background knowledge and recent observations. It accounts for data from three sets of experimental studies of the causal inferring abilities of young children. The model provides a base for further efforts to delineate the intuitive mechanisms of causal inference employed by children and adults, without appealing to inherent principles or mechanisms specialized for causal as opposed to other forms of reasoning.*

### Introduction

What kinds of mental mechanisms must be postulated to explain the causal reasoning abilities of the young child? The issue has been explored by developmental psychologists for many years (Shultz, 1982). As outsiders to this literature, we have been particularly struck by a recent series of studies demonstrating that children as young as 3 to 4 years of age are sensitive to the causal structure of events. They are capable of making reasonable judgments about objects' causal properties after witnessing a few events in which objects appear to play causal roles (Gopnik, Sobel, Schulz & Glymour, 2001; Gopnik, Glymour, Sobel, Schulz, Kushnir & Danks, 2004; Schulz & Gopnik, 2004; Sobel, Tenenbaum & Gopnik, 2004). Young children are also able to categorize objects on the basis of their causal properties (Gopnik & Sobel, 2000). Such findings run counter to the Piagetian view that young children are pre-causal (e.g. Piaget, 1930), and create interesting new avenues for exploring key developmental questions. Here we consider whether it is necessary to invoke a specialized mental structure or module, specifically tuned to representation of causal relations (Gopnik *et al.*, 2004), or whether domain-

general mechanisms of knowledge and cognitive skill acquisition might be sufficient.

#### *Basic findings with the blicket detector*

The findings we will consider arise in studies using the 'blicket detector' – a device that flashes and makes music when certain objects ('blickets') are placed upon it. In the studies considered here, children's categorizations of objects as blickets or non-blickets are assessed, after they observe several events, in the course of which two target objects, when placed on the detector either together or one at a time, either activate the detector or do not (see Tables 1 and 2).

At the outset of these experiments, children are generally shown that some objects activate the detector (and so are blickets) and some objects do not, and appearance cues are carefully controlled so that nothing other than the outcomes of the objects' participation in blicket detector activation events can be used to infer which objects are blickets and which are not. After initial familiarization, a series of events is then presented involving the two target objects. The experiments establish a number of basic points. In 'one-cause' and 'two-cause'

Address for correspondence: James L. McClelland, Department of Psychology, 344 Jordan Hall, Stanford University, Stanford, CA 94305, USA; e-mail: jlm@psych.stanford.edu

**Table 1** *Causal inference tasks (Gopnik et al., 2001)*

Condition	Training sequence						Blicket?	
	Stage 1	Num.	Stage 2	Num.	Stage 3	Num.	A	B
One cause	A+	1	B-	1	AB+	2	1	0
Two cause	A+	3	B+	2	B-	1	1	1

*Note:* Event representations use capital letters (A, B) for objects placed on the blicket detector followed by + and - to denote whether the detector activated in these circumstances or not. 'Num.' columns specify number of presentations of each indicated event. 'Blicket?' columns specify normative values of whether the indicated objects should be categorized as blickets or not (1 = yes, 0 = no).

**Table 2** *Retrospective disambiguation tasks (Sobel et al., 2004)*

Condition	Training sequence				Blicket?	
	Stage 1	Num.	Stage 2	Num.	A	B
Screening-off	AB+	2	A-	1	0	1
Backward blocking	AB+	2	A+	1	1	?

*Note:* Events and normative outcomes are represented using the conventions in Table 1. '?' indicates that whether an object is a blicket or not cannot be determined conclusively from the evidence.

tasks (Table 1), equally frequent occurrences of objects A and B and their outcomes (represented as + or - to indicate whether the detector activates) result in differential patterns of object categorization (Gopnik *et al.*, 2001; Sobel *et al.*, 2004), demonstrating that frequency of co-occurrence of an object's placement on the detector with the detector's activation is insufficient to account for children's responses. It appears that children also take into account information arising from the participation of other objects in the same events and in other events. In both the one-cause and two-cause tasks, there is one occurrence of B alone, with no effect (B-). In the one-cause task there is also an occurrence of A with the effect (A+), and two occurrences of A and B together with the effect (AB+). The evidence here is consistent with the simple hypothesis that A is a blicket and B is not, as indicated in the last column of the table. This is treated as the normative result (Gopnik *et al.*, 2001) for this sequence. In the two-cause task, B occurs once alone with no effect (B-), but also occurs alone twice with the effect (B+). A also occurs alone with the effect (A+). For this case the normative result is said to be to treat both objects as blickets.

In 'retrospective disambiguation' tasks, a later event disambiguates an earlier one (Table 2). In the 'screening-off' condition (Sobel *et al.*, 2004), A- after AB+ is taken to be entirely disambiguating, indicating that B must be a blicket. In the 'backward blocking' condition, A+ after AB+ is not strictly speaking a disambiguating event, in that it leaves open the possibility that B might or might not be a blicket. Nevertheless, in experiments it is found

that the A+ event reduces the tendency for B to be categorized as a blicket, relative to a condition in which the AB+ event is presented alone. This effect is modulated by the proportion of previously experienced objects that are blickets (Sobel *et al.*, 2004): children's tendency to call the B object a blicket is relatively high when blickets have been predominant in previous experience, and much lower when blickets have been rare.

#### *Cross domain inference*

Children's causal reasoning abilities have also been investigated across naturalistic domains using counter-intuitive cause-effect relationships, such as a switch having an effect upon a person, and talking having an effect on a machine (Schulz & Gopnik, 2004). In this study, children were shown to be able to make such domain-inappropriate mappings, crafting interventions illustrating their knowledge. Such mappings were shown to (weakly) influence subsequent judgments of similar situations within the same context, so that domain-inappropriate causes were more likely to be considered as possible causal factors where such causes would not usually be considered at all.

## Representing causal knowledge

### *Causal Bayes nets*

According to the 'theory theory' of Gopnik and Wellman (1994), children approach the world as though they are scientists, developing and testing hypotheses akin to those involved in scientific theorizing. Gopnik and colleagues have recently focused on children's causal reasoning, building on the blicket experiments described above and others to suggest that young children develop accurate representations of causal structure using inductive procedures which allow them to infer a specific type of learned representation - a 'causal map' - from observations of and interactions with the environment (Gopnik *et al.*, 2004). The causal Bayes net formalism (Pearl,

2000) is used as a framework in which to represent such causal maps. The formalism is assumed to provide an abstract framework for characterizing causal knowledge independent of specific content (e.g. physical or biological causation) (Gopnik *et al.*, 2004). It is suggested further that children make use of Bayesian inference to infer causal structure within the possibilities provided by the Bayes net representation (Sobel *et al.*, 2004). Bayesian inference within causal Bayes networks is thought to provide a normative theory toward which human judgments should tend. For example, it predicts that prior experience indicating whether blickets are rare or common should modulate the backward blocking effect.

We fully accept the usefulness of Bayesian inference within causal Bayes networks to provide a normative framework within which to describe the outcome of the causal inference process. The finding that children's behavior tends toward patterns that might be expected within that framework is an interesting and important contribution to our picture of children's thinking abilities. However, the authors of these papers go further than this. In Gopnik *et al.* (2004), it is suggested that a cognitive module specialized for causal inference should be attributed to young children. This module is assumed to apply across domains regardless of specific content, whenever causal inference is required, whereas other mechanisms are said to be employed when other forms of reasoning are needed. Furthermore, it is suggested that this tendency may depend on an innate predisposition specific to causal inference.

In response to this, we ask: Is it necessary, or even useful, to suppose that the child actually possesses a distinct cognitive module for causal reasoning? We have several reservations about the proposal. First, how the causal inferencing module is interfaced with domain-specific knowledge relevant to particular types of causation has not been made clear. This is a general problem for modular approaches: For example, a problem that plagued Marr's attempt to offer a modular approach to vision is the problem of combining outputs from several different modules when each provides relevant but inconclusive results (Bülthoff & Yuille, 1996; McClelland, 1996). Furthermore, in the experimental studies to be considered, 3-year-olds are often less likely to follow the normative pattern than 4-year-olds. This suggests causal reasoning abilities may after all be acquired, perhaps by mechanisms that are domain-general. Beyond these two issues, what we see as the central problem is that it may not be possible even in principle to draw a sharp line between those situations in which causal rather than some other form of inference is required. Suppose two objects are connected to each end of a bar that is free to pivot around a central point. When the apparatus is fully visible, one

could see the physical connection provided by the bar, providing a direct basis for inferring that if a person were to move one of the objects the other would also move. Causal reasoning is not thought to be required in this case, since a principle of naive physics (connected objects move together) is sufficient to reason that if one moves then so will the other. Occlusion of the bar, however, would remove the direct evidence of a physical connection, affording the occasion for invoking the content-independent module for causal inference (Gopnik, personal communication, July 2004). However, when the bar is hidden behind a screen with the objects still visible above its edge, it may be possible to infer the existence of a rigid bar linking the two objects based on perceptual information. Such an inference could be triggered, for example, by evidence that the movement of the objects is very tightly coupled as it would be if there were such a bar. Would this then suddenly remove the problem from the causal realm and return it to the physical? To us it seems likely that any such distinction between physical and causal inference will turn out to be both conceptually difficult to maintain and of little psychological relevance.

Based on the above, we eschew commitments to such concepts as 'cause', 'causal inference', 'causal attribution', etc. insofar as they are thought to refer to entities or properties of specific types within some mental taxonomy of ontological categories. We use such phrases ourselves as a convention to denote tendencies to behave in certain ways in certain situations, which we agree with others to label 'causal inference' or 'causal attribution' tasks. For example, a child is said to attribute a certain causal power to a particular object if, after witnessing an event or series of events in which some effect occurs, the child will then use the object in an attempt to produce the effect if asked to attempt to produce it.

In accord with the concerns raised above, we favor an approach that does not require special modules for inferences of particular types but instead relies on a unified method of (implicit) inference informed by contingent relationships observed in events. We argue that the effect of experience in a domain-general system for learning many different kinds of knowledge can give rise to the ability to make causal inference in accordance with Bayesian principles in causal Bayes nets and that such a mechanism can also account for children's behavior in the kinds of causal reasoning tasks discussed above.

## Domain-general learning and inferencing mechanisms

Specifically, we suggest that domain-general learning mechanisms arising within the Parallel Distributed

Processing approach to human cognition (McClelland & Rumelhart, 1986) may provide an alternative approach for addressing the basis of children's cognitive abilities, including their ability to make causal inferences. In such an approach, children's ability to make causal inferences need not be considered separately from other forms of knowledge.

In our approach, previously sketched out in Rogers and McClelland (2004), knowledge about the causal structure of the environment is thought to emerge from experience with objects and the consequences of their participation in events. Objects can be assigned representations that combine their perceptually identifiable properties with information about the outcome of events in which they participate. Representations capturing the causal properties of new objects participating in such events may be derived on-line from observations, using mechanisms previously used for attributing non-causal semantic properties to objects (Rogers & McClelland, 2004).

When an event is witnessed – that is, an event which involves one object interacting with others such that a particular sequence of sub-events occurs – the effects of the experience shape our representations of the participating objects and give rise to acquired expectations about what would happen at a later time when one witnesses other events. For prior learning to generalize to later events, the representation of the objects in the new event must overlap with the representation of relevant objects that have been previously experienced (McClelland & Plaut, 1999; Rogers & McClelland, 2004). Such overlap can be used to support generalization of properties to new objects. A crucial feature of the models that we use is that all types of information, including information about the causal properties of objects, defined here as information about the possible outcomes of events in which such objects participate, can shape the acquired representation we assign mentally to an object. Objects that participate in similar ways in events come to have similar internal representations, allowing causal inference and generalization.

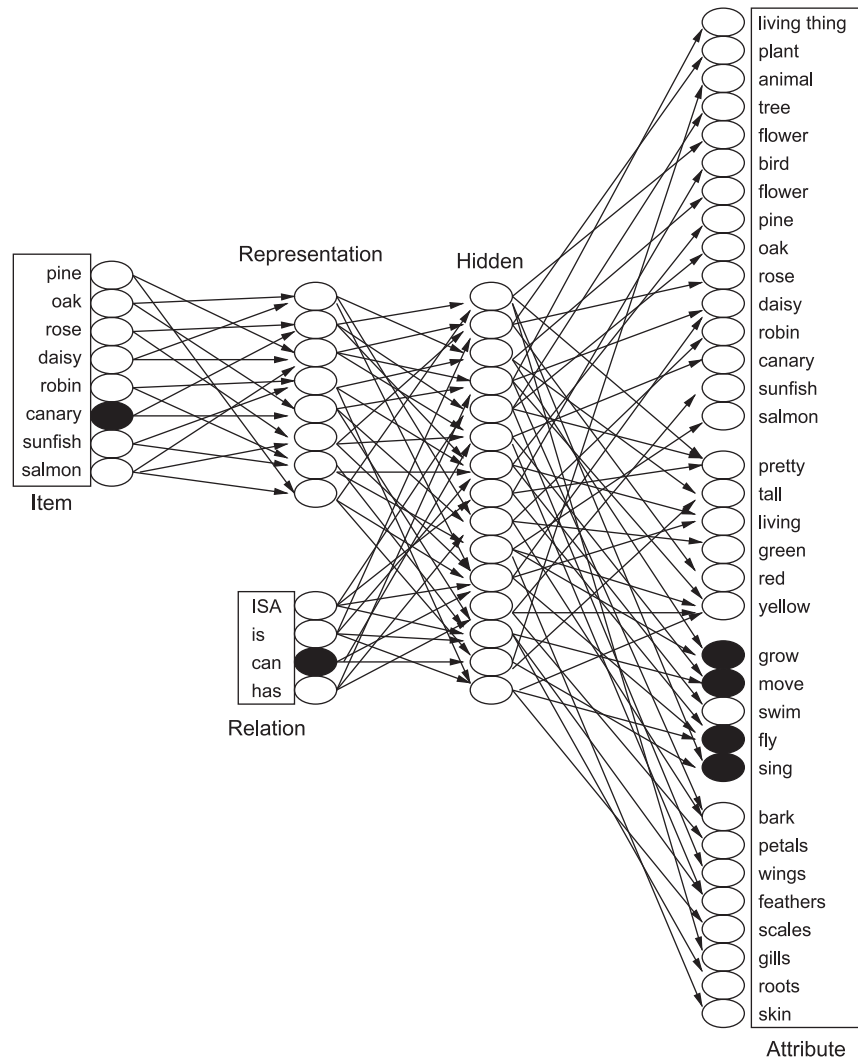
Support for the importance of previous experience may be found in differential patterns of performance between age groups in several of the tasks discussed above. For example, in a backward blocking task (i.e. after exposure to AB+, A+) (Sobel *et al.*, 2004), 3-year-olds were much more likely than 4-year-olds to call the B object a blicket. For the present article, we will not focus on developmental trends, however. Instead we will focus on the surprisingly competent performance of 4-year-olds, whose behavior generally approximates what would be expected of normative causal reasoners in the relevant tasks.

## Foundations of the model

In an early connectionist implementation of semantic memory, Rumelhart and Todd (1993) trained a network (Figure 1) to map from concept-relation pairs of inputs to a list of attributes that stand in the given relation to the given concept. For example, given the input 'robin can' the target output is 'grow', 'move' and 'fly'. This architecture has subsequently been extensively investigated by Rogers and McClelland (2004). The architecture is slightly more complicated than a standard three-layer network. In particular there are two hidden layers, one receiving input from units standing for individual concept terms only and one receiving inputs from this hidden layer and from relation units. The first or representation layer acquires domain-general representations of the concepts based on aggregation of information about their attributes in relations of all types, while the second hidden layer essentially modulates these representations in relation- or context-specific ways to allow them to be used for relation- or context-specific purposes.

While Rumelhart restricted the use of this architecture to learning about relation-specific semantic attributes of objects, Rogers and McClelland viewed the architecture more generally as one that allows for the generation of expected outcomes of situations involving items occurring in specific contexts. They treated item-context pairs as inputs characterizing an item and its participation in a situation, with the task being to learn to predict additional elements of a situation (including successor events) that tend to occur with the item in the given context. For example, we can imagine a blicket on the blicket detector as an item/context pair, and the consequent event of the blicket detector flashing and making music as the successor event. As such we view the architecture as applicable to both semantic and causal attribution.

Both Rumelhart and Todd (1993) and Rogers and McClelland (2004) trained networks of this type using the back-propagation of error, defined as the difference between the network's output (i.e. pattern of activation over the output units) and the externally supplied attribute or outcome information. Networks trained with back-propagation learn to assign distinct internal representations to different items. The representations so assigned are learned internal representations that capture both domain-general and domain-specific patterns of coherent variation among attributes (consequences) of objects observed in different contexts (Rogers & McClelland, 2004). What is learned about one object tends to generalize to some but not all other objects based on the similarity of the learned internal representations. Because the representations differentiate strongly on the basis of type of object (e.g. animate vs. inanimate object) and are



**Figure 1** Structure of the Rumelhart and Todd (1993) network. Conjunctions of item and relation layers elicit a pattern of activity on the attribute units.

sensitized at the second hidden layer to context, generalization exhibits patterns of category (including high-level ‘kind’ categories) and context specificity (Rogers & McClelland, 2004).

A key feature of the Rumelhart network is that the internal representations of objects participate in their use in all contexts. This allows the network’s internal representations to reflect patterns of coherent covariation of properties across contexts. For example, if objects that look like switches have the property of causing state transitions in devices to which they are attached, the network will come to assign an internal representation to such objects that captures both properties at once. We suggest that our semantic and conceptual representations are jointly constrained by experience in this way

and that these representations then are used as the basis for thinking about these objects when they occur in all contexts. The approach does allow, however, that there may be different neural populations in the brain that represent, for example, the predicted and observed appearance properties of an object on the one hand and the predicted and observed consequences of its participation in an event, on the other (e.g. pressing it has the consequence that a device turns on). Thus the causal and appearance properties of the object would depend on partially overlapping and partially distinct portions of the architecture, just as the appearance (IS) and action (CAN) properties of objects in the Rumelhart network depend on partially overlapping and partially distinct portions of that architecture.

The connection weight-based learning in back-propagation networks is very slow – corresponding in our view to gradual change of internal representations over developmental time scales (i.e. months to years). However, once knowledge in the connection weights has been established, it is possible to assign an internal representation to an object based on a single exposure to information about it, and then use this assigned representation to make plausible inferences. Rumelhart and Todd (1993) used this approach to apply knowledge about the typical properties of category members to a novel item based on just one category-relevant proposition. As an example, they examined the generalization of knowledge about typical properties of birds to a new bird – a sparrow – about which the network was only taught ‘sparrow isa bird’. First, they provided a new input unit for the novel concept and froze all weights in the network except for those between this input unit and the representation units, halting any learning in the rest of the network. The process was iterated until weights stopped changing, at which point ‘sparrow isa’ strongly activated the specified target ‘bird’. To achieve this, the network acquired weights from the sparrow input unit to the representation units that made the sparrow input produce an internal representation for the sparrow that closely matched that of other birds. This then supported plausible inferences: Trained with ‘sparrow isa bird’, the network generated appropriate activations of typical bird attributes when tested with ‘sparrow has’, ‘sparrow can’ and ‘sparrow is’.

One can use this approach, not to generate connection weights, but to generate a pattern of activation amounting to an assigned representation for an item. Indeed the back-propagated error signal to the units at the representation layer specifies how their activations should change to produce a representation consistent with the specified target information, and Rogers and McClelland employed this approach in their simulations. This method, called back-propagation to representation, was initially proposed by Miikkulainen and Dyer (1987) to assign representations to words based on their roles in sentences. Rogers and McClelland (2004) made extensive use of this method in their simulation of children’s generalization of object properties, showing that the representations produced by this method allowed the network to use what it had previously learned about different types of items and their attributes in different contexts to produce representations supporting domain-appropriate generalizations. In the present work we will use this approach to assign representations to objects based on what others have called their ‘causal powers’; that is, the tendency for particular outcomes to occur when these objects participate in particular ways in particular contexts.

One view of the above approach is that it simply provides a way of allowing a back-propagation network to simulate an interactive activation process, and we would not want to completely discourage this view. That is, we do not intend to endorse the notion that activation flows only in one direction and error information flows only in the other. Indeed, it has been shown (Hinton & McClelland, 1988; O’Reilly, 1996) that activation signals can carry error information.<sup>1</sup> On the other hand, the use of back-propagation to representation captures more directly than standard activation propagation schemes the idea that we use the mismatch between predicted and observed information to change our internal representations in a way that allows the representation to account better for what we observe. Such an effect is not necessarily the consequence of the propagation of activation information (although it would be in the networks of Hinton & McClelland, 1988, and O’Reilly, 1996).

One feature of many of the experiments we will consider is that they involve the child observing several events, from which he or she must form a representation of the causal powers of participating objects. Some events are ambiguous – two potential causes appear, and if a consequence ensues it is not possible to determine which of the objects should be given responsibility for it. Crucially, what the child needs to do is to assign causal powers to each of the objects participating in these events so as to be able to account for all of the events. Specifically, in the retrospective screening-off task, the child witnesses AB+, A–, and both 3- and 4-year-olds reliably report that both A and B are blickets. The causal power of A can be determined by the second event alone, but the causal power of B can only be assigned by considering both events. Assuming blickethood is all-or-nothing, and assuming that the effect occurs if one or more of the objects on the detector is a blicket, it is possible to infer from the combination of the two events that B must be a blicket. How might children actually use the two events to determine that B is a blicket?

One way children could do this would be to use the back-propagation representation procedure to find representations for each participating object that work simultaneously to account for all of the events presented. This can be achieved by interleaved presentation of each of the two situations, gradually adjusting the representations of each item on each presentation until they stop changing. Indeed, Rogers and McClelland used back-propagation to representation in just this way to assign

<sup>1</sup> More accurately, error information is carried by the difference between activation signals returning from a network’s output units over which it predicts outcome information, before vs. after the arrival of outcome information that affects the activations of the units.

a representation of a novel item based on information from several different contexts. For example, they could constrain a representation with the information 'X has roots' and 'X is big' by interleaving presentation of each and gradually adjusting the representation iteratively until a representation satisfying both pieces of information was found. The resulting representation was close to that of a typical tree, allowing imputation of other tree-like properties when the representation was reinstated in other contexts. We suggest this procedure as a form of intuitive reasoning, based on knowledge stored in connection weights, and giving rise to distributed representations sensitive to the implications of two or more distinct observations. The psychological plausibility of the procedure is a source of concern, however – the rapid mental interleaving of events might take time, and could place high demands on control processes. We will return to this issue in the general discussion.

A final issue for which we rely on earlier work is the question of how the representation of an item once formed is retained for later use. Here we rely on the complementary learning systems theory of McClelland, McNaughton and O'Reilly (1995): The theory assumes that experience gradually shapes the connections among units in a slow-learning, 'neocortical' learning system, providing the background knowledge individuals bring with them to a new situation. This system is complemented by a fast-learning system thought to depend on the hippocampus and related structures in the medial temporal lobes. In the present context, the second, fast-learning system learns the association between perceptually distinguishing properties of an object and the internal representation assigned to it on the basis of its participation in a series of events. The association is stored in the form of connection weights that can be accessed later when the object is presented again in a test situation. In this way the representation of an object's causal properties, created during an exposure phase, can be reinstated when the object reappears later. This idea was previously discussed in Rogers and McClelland (2004) for use in semantic tasks. It provides a general mechanism for 'fast mapping' of a representation of an object based on observed properties from a single learning session (Bloom, 2000; Carey, 1978), of which its use here is an instance.

### Aims of the reported simulations

The aim of these simulations is to instantiate the perspective on causal inference described above in the form of the simplest possible computational implementation with the goal of making the mechanisms at work as

explicit and transparent as possible. With the implementation we will demonstrate how the ideas described above, previously developed in the context of acquisition and use of non-causal semantic and conceptual knowledge, can be applied without major alteration to tasks typically viewed as requiring causal reasoning.

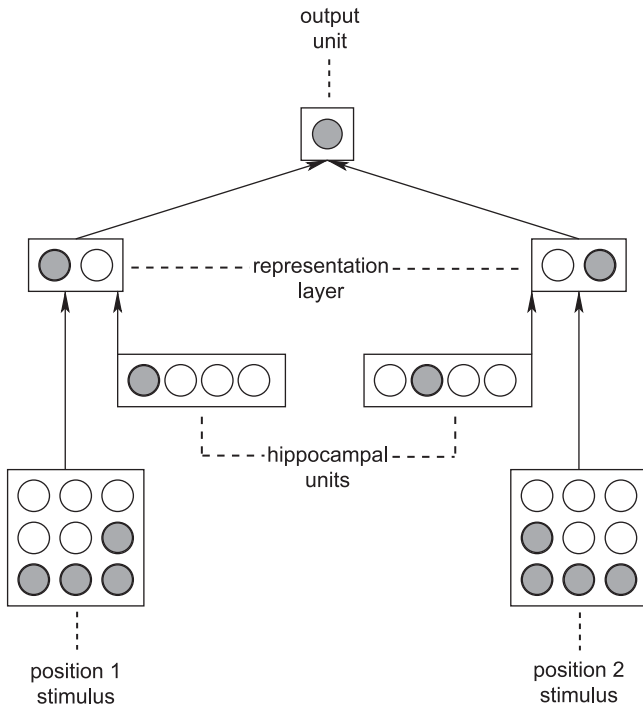
The model is intended to simulate findings from Gopnik *et al.* (2001), Sobel *et al.* (2004) and Schulz and Gopnik (2004), to show how:

1. A representation can be assigned to an object based on its experienced causal powers, that is, on the outcomes of events in which it and other objects participate.
2. Inferences about the causal powers of novel objects can be made based on shared properties with other objects.
3. Frequency of occurrence of causal powers can influence causal attribution.
4. Previous experience in particular kinds of situations can:
  - (a) Influence initial attributions
  - (b) Be overridden by subsequent observations.
5. Experience within a context can bias the types of causes which will subsequently be considered in that context.

### Model structure

The basic structure of the model is illustrated in Figure 2. The model contains a 'cortical' network together with a complementary 'hippocampal' learning system. We describe the cortical network first. This network is intended to have the minimal architecture necessary to provide an analog of the relevant background knowledge children bring to theblicket experiments. It contains two stimulus arrays, two hidden or representation layers, and a single output unit. Two stimulus input arrays are used so that objects can be presented to the network either one or two at a time. The task facing the network is to learn from situations involving one or two inputs whether a particular outcome, e.g. activation of theblicket detector, should be predicted. Since we are concerned only with this single binary outcome only a single output unit is needed.

Each object presented to the network is represented as a pattern of activation over one of the stimulus arrays. These patterns are intended as proxies for the derived internal representations formed through experiences with objects in different contexts, which arise from the presentation of an item over the representation layer in the Rumelhart and Todd network described above. For concreteness, units in these patterns may be thought of as coding for features of the objects that may bear some



**Figure 2** The basic structure of the model used in simulation 1. Stimulus and hippocampal units are input layers; representation layer is a 'hidden' layer. Cortical training uses the stimulus, representation, output pathway; hippocampal training involves weight matrix between hippocampal and representation layers. Arrows denote fully interconnected weight matrices.

predictive relationship to the object's causal powers. Importantly the features need not be perceptual primitives such as 'blue', 'square', etc. They can also capture learned conceptual distinctions such as 'human', 'animal', 'electrical device', 'switch', etc. that we take as important for representing covarying clusters of properties observable in different contexts. Each position-specific array is fully connected to its corresponding representation layer, consisting of two units, and the units in both representation layers are in turn connected to the output unit. The hidden layer allows the network to assign a learned internal representation to each stimulus reflecting its tendency to cause the specific outcome represented by the single output unit. For our later simulations more than one causal outcome is considered, so additional output units and additional input units to represent the causal context are used. This will then allow the representations formed at the representation layer to be context-sensitive. In the present simulation we consider only a single context so these context units are unnecessary.

The architecture provides a way to represent at the same time two objects that occupy equivalent roles in an event,

namely that of being placed on the blicket detector. To allow generalization across positions, the weights are shared over the two position-specific input-to-representation and representation-to-output matrices, so that they are kept identical. The combined stimulus-to-representation weight changes calculated for each matrix separately are applied to both matrices every time the connection weights are changed. Changes to the connections between each representation layer and the output layer were combined in the same way. Note that the need to generalize knowledge over positions is a general one that has been faced before in connectionist networks, and several different implementations have been proposed (McClelland, 1986; Mozer, 1987). The weight-sharing approach used here was first used by Rumelhart, Hinton and Williams (1986a) in a model of position-invariant object recognition.

In addition to the cortical system, there is also a separate hippocampal learning system, consisting of the hippocampal units and weights from these units to the cortical representation units. As with the cortical system, two copies of these units and connections are provided, and weight sharing is used to allow hippocampal learning about objects to generalize across positions. Corresponding to these two learning systems, there are two major phases in the training of the network. The cortical learning phase provides the basic background knowledge acquired gradually through prior experience, while the second 'hippocampal' phase uses the background knowledge to interpret experiences during the experiment, then stores them in connections to and from a separate pool of hippocampal units. For the reported simulations, training in both phases continues until representations stop changing, which eliminates the number of training cycles as a parameter of the simulations.

It is possible that an individual with extensive hippocampal damage would be able to perform successfully in tasks like the blicket detector task, where there is little or no delay between the exposure and test phases of the experiment; while hippocampal learning would certainly be required to sustain a delay, it may be possible to rely on maintained activation in working memory. We have not implemented a system for doing this, relying on the hippocampal system, which would be required if there were a delay, to play the role of the working memory representation in this case.

### Cortical pre-training to produce relevant background knowledge

Young children do not have prior experience with blicket detectors *per se*, but we assume that the children in these experiments have experiences involving objects producing

outcomes similar enough to those occurring in the blinket experiments for there to be a base of prior knowledge on which they build in the experiments. Pre-training of the cortical portion of the network provides a proxy for this implicit background knowledge children are assumed to bring to the experiment.

Weights are trained using the back-propagation algorithm (Rumelhart *et al.*, 1986a), with weight sharing as described above. Training occurs through a series of epochs, during each of which each member of a set of background experiences is presented in succession. Weight adjustments are cumulated over each experience in the epoch. These cumulated adjustments are added to the connection weights in the network at the end of the epoch, and the weights are subjected to weight decay at the same time.

The range of stimuli trained is intended as a simplified proxy for a broad range of experiences with objects possessing various properties, including causal properties. The model extracts generalizations from these experiences – specifically, it learns which object features tend to predict their causal power to activate the blinket detector. These generalizations are encoded in the weights learned, causing the network to have different expectations of the causal powers of different test objects based on their particular features.

Weight vectors from hippocampal units to representation units were frozen during cortical pre-training so that no learning occurred on these. In the theory, of course, hippocampal as well as cortical learning occurs all the time, but the theory also holds that hippocampal traces decay more rapidly and are both item- and context-specific, such that the knowledge acquired gradually in the cortical learning system will be the primary (systematic) basis of prior knowledge children bring to the experimental situation from prior experience.

### Hippocampal learning during experimental exposure

Weights in the cortical system change very slowly on a time scale of the standard psychology experiment, and are thus treated as fixed on that time scale for simplicity. Intra-experimental learning depends on the much more rapid learning afforded by the hippocampal system (McClelland *et al.*, 1995). Hippocampal learning is therefore used to store experiences acquired during the experimental training period, when objects are presented in conjunction or alone and shown to have some effect or no effect.

Each stimulus item used during experimental exposure is assigned to a single specific hippocampal unit, on

which experimental learning relies. This practice accords with the assumption that the hippocampal representations are sparse and highly conjunctive, so that even very similar inputs have little overlap in their hippocampal representations (McNaughton & Morris, 1987; O'Reilly & McClelland, 1994). It is necessary to assume in fact that these representations can reliably be associated with the appropriate object even when it is perceptually indistinguishable from other objects being used; this would require keeping track of its individual identity through space, as it is moved on and off the blinket detector. As in the cortical networks, weight sharing is used to allow presentation of an experimental stimulus in either input array to activate a representation of the object's causal powers over the corresponding representation units.

The learning during this stage involves the use of the back-propagation algorithm to assign representations for the experimental stimuli (corresponding to the A and B items in the experiments), and for these to be used to set the weights from the hippocampal unit to the representation units, so that activation of the hippocampal unit can reinstate the item's acquired representation (see Figure 2). All other weights in the network are frozen during this phase of learning. The goal is to find representations for both objects that, when combined with the knowledge already stored in the cortical system, allows the network to account for the causal outcome observed in each of the events occurring in the exposure phase of the experiment. This is achieved by iterating over the set of events used in the exposure phase of the experiment, presenting them in interleaved fashion and adjusting the representations (and hippocampal weights) incrementally until equilibrium is reached, as previously discussed. In each iteration, each of the experiences from the exposure phase of the experiment is presented in succession. Whenever a particular object is presented in either input channel, its assigned hippocampal unit is also activated in the corresponding channel. Error is back-propagated to the representation layer, where the resulting signal is used to derive adjustments to the links to this representation layer from the corresponding hippocampal unit. These weight adjustments are cumulated over all of the exposure-phase events, then added to the weights after all of the experiences have been presented, which are also subjected to weight decay at the same time. The iteration continues until the weights stop changing.

### Simulations

Simulations were carried out using the Lens neural network simulator (Rohde, 1999) with enhancements added

to implement weight sharing as described above.<sup>2</sup> The default learning algorithm, which normalizes the length of the weight error derivative vector at the end of each epoch, was used with learning rate of 0.1 and momentum of 0.9. These are the default values of the simulator, and were fixed throughout the simulations.

The only free parameter of the simulations was the weight-decay parameter. Weight decay imposes a pressure against allowing connection weights to grow large, causing them to regress slightly toward 0. The value of the weight-decay parameter determines the strength of this pressure. We report results based on a very small value of the weight-decay parameter, which allows weights to grow very large, and thereby allows the network to approach within about 1% the extremal activation values of 0 and 1 in many cases.

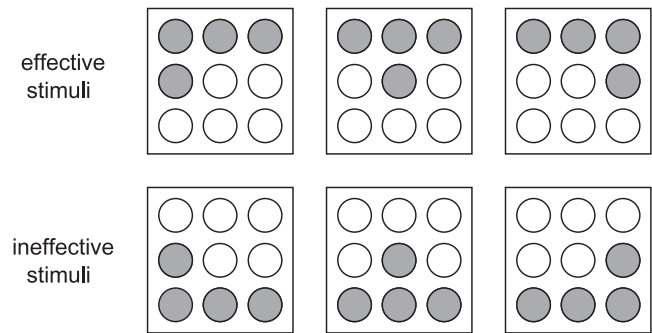
#### *Simulation 1: Deriving causal properties of objects from observing the outcomes of their participation in events*

The aim of this simulation was to perform the ‘one-cause’ and ‘two-cause’ tasks used in Gopnik *et al.* (2001) and the ‘indirect screening-off’ task used by Sobel *et al.* (2004) in order to illustrate how the model derives differential representations dependent upon the experienced causal powers of each object as well as those of other objects. We consider the closely related ‘backward blocking’ paradigm also used by Sobel *et al.* (2004) in simulation 2, where we also consider how the model can take into account prior base-rate information. The structure of the network, shown in Figure 2, has already been discussed.

#### Cortical pre-training

For the cortical training, two types of stimulus patterns were used: ‘effective’ and ‘ineffective’ (see Figure 3). Ineffective patterns were those which, when presented alone, had a target output of 0. Effective patterns were those which, when presented alone or in conjunction with ineffective patterns, had a target output of 1. Each stimulus had an associated input pattern over the three by three input array. Effective stimuli activated all the units in the first row and one unit in the second; ineffective stimuli activated all units in the third row and one in the second. Cortical training involved presenting 27 patterns per epoch; the three effective patterns three times each, the three ineffective three times each, and one presentation of each possible combination of effective with ineffective patterns (a total of nine combinations).

<sup>2</sup> This version of the simulator plus all of the files used to train and test the networks are available by contacting the first author.



**Figure 3** The stimuli used in the first training stage. Effective stimuli represent objects which have causal powers and are always paired with a positive output in the model; ineffective stimuli represent objects which have no causal powers, and have no influence over the output.

The effect of the pre-training was to establish input-to-representation weights producing two different representations, one shared by the effective stimuli and one shared by the ineffective stimuli. The pre-training also established representation-to-output weights which operated in conjunction with the input-to-representation weights so that the presentation of an effective stimulus in either position produced a strong activation (close to 1.0) of the output unit and produced an activation close to 0 otherwise.

#### Exposure phase of experiment

For the exposure phase of the experiment, the input patterns used were ambiguous; each one had a single unit on each of the three rows. When tested individually such patterns produced an intermediate activation of the output unit (between 0.34 and 0.35), representing uncertainty on the network’s part about the causal status of the pattern. We also repeated the simulation using patterns with only a single unit on the ‘ineffective’ row. When tested individually, such patterns produced a very low activation of the output unit, i.e. before exposure to the blicket detector the network viewed them as ineffective. As the results were essentially the same in both cases we report only the results from the former testing condition.

Table 3 shows schematically the set of experiences used in the simulation of each of the three experimental conditions considered. In the ‘screening-off’ task, called ‘indirect screening-off’ in Sobel *et al.* (2004), two novel patterns in combination activated the detector, twice. Then one of the patterns was presented as ineffective. The ‘one-cause’ and ‘two-cause’ tasks (refer to Table 1) were identical to those in Gopnik *et al.* (2001). In the

**Table 3** Training events used in simulation 1

Screening-off			One-cause			Two-cause		
P1	P2	O	P1	P2	O	P1	P2	O
n1	n2	1	n1		1	n1		1
n1	n2	1	n2		0	n1		1
n1		0	n1	n2	1	n1		1
			n1	n2	1	n2		1
						n2		1
						n2		0

Note: P1 is stimulus position 1. P2 is stimulus position 2. O is target output. n1 is novel pattern 1. n2 is novel pattern 2.

‘one-cause’ task, individual presentations of one effective and one ineffective novel pattern were followed by presentation of a combination of both patterns, which activated the detector twice in succession. The ‘two-cause’ task involved presentation of an effective novel pattern three times, followed by presentation of an object which appeared to be effective twice, and ineffective once. Note that in all three conditions, one object always occurs with the effect, and is called the *100% effective object* while the other object always occurs twice with the effect and once without, and is called the *67% effective object*. For each condition, one of the three exposure phase stimuli previously mentioned was assigned to the role of n1 and one to the role of n2.

Note that in the experiments, different subjects were used in each condition. Correspondingly, a fresh copy of the network, with the pre-trained cortical connection weights in place, was used in the simulation of each of the three conditions. Training occurred according to the procedure described above.

### Test phase

To test a network after the exposure phase, we simply present the two patterns used in training one at a time, in one of the two input slots, and record the resulting activation of the output unit. Presentation of the pattern involves turning on both the pattern itself and the corresponding hippocampal unit. Taken together, these have the effect of reinstating the representation of the pattern obtained at the end of the experimental exposure phase described above.

### Results and discussion

The simulation results are shown in Table 4 along with the pattern expected from normative performance and along with the results of the relevant behavioral experiments. Numbers reported are, for children, the proportion of ‘Yes’ responses to the question ‘Is it a blicket?’,

**Table 4** Categorization of different blocks as blickets

Source	Screening-off task		One-cause task		Two-cause task	
	100%	67%	100%	67%	100%	67%
Normative	1.00	0.00	1.00	0.00	1.00	?
Gopnik <i>et al.</i> (2001)	–	–	0.91	0.16	0.97	0.78
Sobel <i>et al.</i> (2004)	0.99	0.04	–	–	–	–
Model	0.99	0.01	0.99	0.01	0.99	0.67

Note: Experimental results are proportions of children’s responses to the different blocks in the different conditions. Experimental proportions are based on two trials with each of 16 children per condition. Model results are activation values for the network’s output unit.

or, for the simulation, the strength of activation of the output unit when presented with the test input as probe. Results from Gopnik *et al.* (2001) are from 4-year-olds in their Experiment 1. Three-year-olds were much more likely than 4-year-olds to call 67% effective objects blickets in the one-cause task (66% yes for 3-year-olds, 16% yes as shown for 4-year-olds), and slightly more likely to call 67% effective objects blickets in the two-cause task (85% yes for 3-year-olds vs. 78% yes for 4-year-olds). Results from Sobel *et al.* (2004) are combined over the 3- and 4-year-old age groups in Experiment 1 and the single (4-year-old) group in Experiment 2 (the 4-year-old group in Experiment 1 showed perfect performance, while low rates of errors occurred in the other two groups).

There are many ways to relate the model’s activations to the proportion of responses made by young children. One could specify, for example, that whenever the activation of the output unit exceeds some threshold (say, 0.5) the model is taken to have attributed the causal power to the object presented. In fact, however, the activations assigned by the model actually approximate an estimate of the probability that an object has a particular causal power. If one thought that children’s individual responses were random samples from a process that also matches these probabilities, then the network’s activation values could be seen as representing a prediction of the expected proportion of yes responses for each condition. The choices made by adults, children, and non-human animals often conform to such a probability matching rule, and we therefore see it as interesting to compare the network activations to the children’s response probabilities. We follow this course with the proviso that other rules for relating activation to response probabilities are possible. For this reason we do not speak of the model’s output as predicted response probabilities, referring to them simply as activation values, whose correspondence to response probabilities is worth considering.

With the above in mind, it is apparent that the model conforms well to the normative pattern and captures in

idealized form the basic structure of the experimental data. The activation produced by the network matches the normative patterns for the screening-off and one-cause task to within the limits imposed by the weight decay, and represents cleaner performance than exhibited by children, especially for those objects that normatively should not be called blickets (67% stimuli in the screening-off and one-cause experiments).

There is one place where the normatively correct response is unclear: the case in which a single object, placed on the detector three times, twice appears to cause flashing and music and once does not. In this case 4-year-old children are fairly likely to treat the object as a blicket (and 3-year-olds are even more likely to do so). Exactly why children exhibit this pattern is not clear. It is clear, however, that the network is setting its output to match the probability that the B object is associated with the positive outcome. This is a well-known consequence of using the back-propagation learning algorithm when the output is not deterministic; setting the output activation equal to the probability that the output unit should be on for a given input results in the smallest summed value of the output error across the different occurrences of that input, and thus is the value toward which the error correcting learning procedure will tend to converge. While the 4-year-old children called such objects blickets somewhat more often than would be predicted from probability matching, the actual value falls within the 95% confidence interval for probability matching, based on the small number of responses per condition, so that the idea that the 4-year-old children are probability matching cannot be rejected.

In summary, a network using a domain-general architecture derived from research in non-causal domains can behave in accordance with normative rules of causal inference, producing more idealized performance than young children. Children's behavior can be related to that of the model by assuming that children derive an estimate of the probability that an object is a blicket and then probabilistically choose the blicket response at a rate equal to this probability estimate. Children's behavior does not exactly match this ideal, especially for those objects that should not be called blickets, but approximates it reasonably closely.

It may be useful to make clear note of the fact that the simulation does not produce correct results if the learning is carried out pattern by pattern, in the order of presentation to children. In the case of screening-off, for example, the network would assign an intermediate representation to both A and B after training with just the first pattern, and this would remain unchanged after the second pattern. The presentation of the third pattern would result in a change in the representation of pattern

A, as required, but would leave pattern B unchanged, since its absence from the input in this case would provide no basis for adjusting its representation. The very same problem would arise if the standard Rescorla-Wagner learning rule were used in the usual sequential order of pattern presentation, and for essentially the same reason. Indeed, backward blocking was first introduced in conditioning (not causal reasoning) paradigms and has provided a strong challenge to that model.<sup>3</sup> This is consistent, of course, with the logical point made in the introduction, that it is essential to allow the representation of both objects to be affected by the information provided in both events. Interleaved presentation allows for this, but sequential presentation does not.

### *Simulation 2: Backward blocking and incorporating knowledge of priors*

The aim of our second simulation was to address the phenomenon of 'backward blocking', a form of retrospective disambiguation. As previously discussed, in the backward blocking task the child is first shown blocks A and B together activating the blicket detector (AB+), and then is shown A activating the detector alone (A+). The effect of the presentation of A alone is assessed by comparing children's tendency to call B a blicket after witnessing only the AB+ event to their tendency to call B a blicket after witnessing both the AB+ event and the A+ event. We also wish to address the fact that 4-year-old children are able to incorporate prior probabilities into such decisions as shown in Experiment 4 of Sobel *et al.* (2004), so that their tendency to call B a blicket is greater when the proportion of blickets previously seen is high than when it is low.

### The experiment

We focus on Experiment 4 of Sobel *et al.* (2004), in which children were shown, in a 'pre-exposure' phase, that blickets were either rare or common. Throughout the experiment, all the blocks were identical in shape and color, and the child was informed at the outset that the task was to figure out which of these were blickets (i.e. blocks that activate the blicket detector) and which were not. In

<sup>3</sup> The challenge to the Rescorla-Wagner model has been addressed in a variety of classical conditioning models (Krushke & Blair, 2000; Schmajuk & Larrauri, 2006; Van Hamme & Wasserman, 1994). These models explain the backward blocking without using explicit reconsideration of earlier events, and the mechanisms they postulate may be at work in some backward blocking situations. We focus on other mechanisms which we suggest may be at work when people are explicitly considering what they should infer from a few specific events they could be holding in mind simultaneously.

the rare condition, only two of 12 blocks the child saw in the pre-exposure phase were blickets, while in the common condition, 10 of 12 were blickets. In both conditions, the first pair of blocks was demonstrated activating the detector together, then children were shown that one member of the pair activated the detector by itself, and that the other did not. The children were carefully instructed that blickets activate the detector either when they occur alone or when they occur together with another block that might or might not be a blicket. The block that activated the detector when placed on the detector alone was placed in a box labeled 'blickets', and the one that did not was placed in a box labeled 'not blickets'.

Five more pairs were then shown. Within each pair, the blocks were placed on the detector one at a time and if a block activated the detector, the child placed it in the box labeled 'blickets'; if it did not, the child placed it in the box labeled 'not blickets'. In the common condition, nine of these 10 blocks were blickets; in the rare condition, only one was a blicket. After this phase, children's attention was explicitly called to the fact that most of the blocks were blickets in the common condition or that few were blickets in the rare condition. Also, the blocks from the pre-exposure phase remained in the boxes as an explicit reminder of the number of blickets and non-blickets previously seen. In the test phase that immediately followed, each child participated in a backward blocking trial, in which the child observed two blocks in an AB+ event (both on detector which activates) followed by the A+ event (one on detector which activates) and was then asked to place each of the blocks in the appropriate box (blicket or not blicket). The child then participated in a baseline trial in which two new blocks were shown activating the detector together (the AB+ event). Again the child was asked to place each block in the appropriate box. In both test trials children were told to guess if uncertain. For 4-year-old children, the probability of placing the B block in the blicket box was much higher in the common than the rare condition (see Table 6). Three-year-old children had a tendency to treat the B block as a blicket in both conditions, with only a slight trend to treat it as a blicket more often in the common condition (88% vs. 81%).

#### Model assumptions

The model is based on the same cortical network with the same cortical pre-training as before. To address the pre-exposure phase, we implemented a simulation that captures the idea that the child forms and maintains a single representation capturing the probability that an experienced block is a blicket, based on the set of objects presented during the pre-exposure phase of the experi-

ment. This representation was formed by adjusting it in response to observations of the consequences of placing each block singly on the detector. The set of patterns used for this consisted of six input patterns, each involving a single unit active in each row of the input. For the common condition, five of the six were paired with a target output value of 1, and one was paired with the target output value of 0, while for the rare condition, five were paired with a target of 0, and one with a target of 1. (Pairing 2/12 or 10/12 patterns with a target of 1 would have produced the same results.) As in the exposure phases of earlier simulations, the set of patterns was presented repeatedly in an interleaved fashion until the representation stabilized.

A single unit which can be thought of as standing for the typical blicket was assigned connection weights to the units in the representation layer, so that when this unit was active it would adjust the representations to lead it to predict an activation for the output unit in line with the proportion of blickets experienced in the pre-exposure phase. For convenience of implementation we used a unit in the hippocampal layer of the network, but we suggest that this unit (and its connections to the representation layer) should best be viewed as a proxy for perceptually available information that would have produced the same effect. This is because the child's attention was called explicitly to the two boxes labeled blickets and non-blickets at the beginning of the final exposure phase of the experiment, and the high or low proportion of blickets (depending on the pre-exposure condition) was noted. Furthermore, the labeled boxes containing the blocks that had been determined to be blickets and non-blickets remained visible throughout the rest of the experiment. Thus the information that many or few blocks are blickets may be perceptually maintained rather than hippocampus dependent. In any case this unit remained active and the weights from this unit to the representation layer remained fixed for the remaining phases of the experiment, in keeping with the fact that the perceptual support for the proportion of blickets stayed constant and remained available in the actual experiment.

For the final exposure phase of the experiment, new hippocampal units were used to encode information about the final pair of blocks, using patterns shown in Table 5. The 'typical blicket' unit discussed above was active throughout this phase, allowing the representation of each block to inherit what the network had learned about similar blocks drawn from the same pool during the pre-exposure phase of the experiment. Otherwise the final exposure phase followed the same procedures used for the experimental exposure phase in simulation 1. The backward blocking and baseline conditions shown in the table were run independently, restarting the network

**Table 5** Patterns used during hippocampal training in simulation 2

Backward blocking			Baseline		
P1	P2	O	P1	P2	O
n1	n2	1	n1	n2	1
n1	n2	1			
n1		1			

Note: P1 is stimulus position 1. P2 is stimulus position 2. O is target output. n1 is novel pattern 1. n2 is novel pattern 2.

**Table 6** Bliкет responses to blocks occurring in various situations

Condition	Source	Initial	Task	
			B. blocking	Baseline
Rare	Sobel <i>et al.</i> (2004)	–	0.25	0.78
	Model	0.17	0.22	0.67
	Estimated	0.17	0.23	0.55
Common	Sobel <i>et al.</i> (2004)	–	0.81	0.97
	Model	0.83	0.84	0.86
	Estimated	0.83	0.85	0.86

Note: Proportion of bliкет responses for children in Sobel *et al.* (2004); activation of the output unit for the model, and estimated bliкет probability based on assumptions given in text.

each time from its state at the end of the pre-exposure phase of the simulation.

The two blocks from the final pair were then tested during the test phase. As in the final exposure phase, the ‘typical bliкет’ unit remained active, allowing it to influence the network’s attributions of causal efficacy to both members of the final pair of blocks.

## Results and discussion

The empirical findings are compared with the simulation results in Table 6. Also included are the outputs generated by the network for the B block prior to either the backward blocking or baseline presentations.

It is evident that the network was able to learn the appropriate expectation for the common and rare conditions. Specifically, following pre-exposure the network sets the activation of the output unit to .17 when a new block is presented in the rare condition, in which one of six pre-exposure objects are bliquets, and sets the activation to .83 in the common condition, where five of six objects are bliquets. As previously discussed, this occurs because the minimum error in the network’s output is reached when the activation of the output unit for a given situation equals the probability that the output is 1 in that situation in its experience.

It is also evident that the network is able to use this knowledge to influence its responses to the B block after both the backward blocking and baseline experiences. Similar to the children in Sobel *et al.* (2004), the network is generally more likely to treat a block as a bliкет in the common than the rare condition, and indeed for the backward blocking condition, there is close numerical correspondence between the network’s assigned activation value and the children’s probability of calling the object a bliкет. Interestingly, after the baseline exposure (AB+), children called the B block a bliкет more frequently in both the rare and the common conditions than the network did.

The difference between the network and the children prompts us to ask: What probability should the child, or the network, assign to the B block in the backward blocking condition, and to both blocks in the baseline condition? We derived estimates based on the following assumptions.

1. For all exposure events, the experimenter is believed to be selecting blocks at random from a large enough pool that the probability  $p$  of drawing a bliкет on each draw has a fixed value. This value, however, is not known, and must therefore be estimated.
2. After each new experience in which a bliкет is singly placed on the detector (or in which two blocks are placed on the detector and it does not activate, demonstrating that neither is a bliкет), a counter for the total number of blocks is incremented, and a counter for the number that activate the detector is incremented if appropriate. An estimate ( $\rho$ ) of the bliкет probability is made by taking the ratio of effective ( $e$ ) to total blocks ( $t$ ). This ratio is known to be the optimal estimate of  $p$ , assuming a null prior.
3. After an experience in which two blocks are placed on the detector and the detector activates, the  $e$  and  $t$  counters are not updated, since the event is ambiguous. However, when required in the baseline condition, an estimate that either of the blocks is a bliкет is made, applying the following reasoning: Since the detector activated, there are three possibilities: (A+,B–); (A–,B+); and (A+,B+). The true probability of the first or the second is  $p(1-p)$ ; the true probability of the third is  $p^2$ ; the true probability that either block is a bliкет is therefore

$$(p(1-p) + p^2)/(2p(1-p) + p^2) = 1/(2(1-p) + p)$$

This result is used to calculate an estimated probability that either block is a bliкет, substituting the value of  $\rho$  from the end of the pre-exposure phase into the above equation in place of  $p$ .

4. In the backward blocking condition, the estimate of  $\rho$  is updated after observing that the A block is a bliкет.

The implications of the above are that:

1. The experienced ratio of effective  $e$  to total blocks  $t$  from the pre-exposure phase will serve as the initial, post-exposure estimate, i.e. this estimate would simply be  $e/t$ .
2. In the 'baseline condition', the estimated probability that either block is a blicket is  $1/(e/t + 2(1 - (e/t)))$ .
3. After experiencing A+ in the backward blocking condition, the estimate of the probability that the B block is a blicket will change from  $e/t$  to  $(e + 1)/(t + 1)$ .

The resulting values based on the actual numbers  $e$  and  $t$  from the common and rare conditions are shown on the line labeled 'estimated' in the table. It will be seen that both the model and the children behave in accord with the estimated probability value in the backward blocking condition, but that the children, and to a lesser extent the model, over-estimate that B is a blicket in the baseline condition. While neither the model nor the children can be seen as matching the indicated estimates perfectly, the model comes considerably closer than the children.

The network's behavior fails to conform perfectly with the estimated model because the method used for combining information across two input positions is not exactly equivalent to the logical 'or' operation, which is, given the logic of blicket activation, the correct way to combine the two sources of information. It might then be argued that the model does not precisely reflect the causal logic of this situation. While we concede this we would also note that by the same token, the behavior of children also fails to reflect this causal logic precisely as well. We will return to this issue in the general discussion.

### Simulation 3: Interplay of prior 'domain-specific' expectations and experimental observations

#### Aim

The aim of this simulation was to show how the model, like children, can exhibit a bias to choose a contextually

appropriate action over a contextually inappropriate one based on pre-experimental experience, while still allowing this tendency to be overridden by experience within an experimental session. Once again no special mechanisms for causal inference are invoked; rather we rely again on the same domain-general mechanisms we have relied on throughout our simulations.

#### Data

In a series of studies, Schulz and Gopnik (2004) have shown that context-specific inferences – that is, inferences involving specific types of actions that typically cause specific types of effects – are based initially on context-specific expectations, but these can be overcome by evidence that goes against these expectations. In their Experiment 4, a task they place in the 'physical domain', involved determining which of three possible causes (attaching either of two magnetic buttons or speaking to it) could influence a noise-making machine. The available appropriate action was to attach a magnetic button. The inappropriate action was to speak to the box. A task in their 'psychological domain' involved a person giggling and three possible actions; showing either of two silly faces stuck to lollipop sticks, or flipping a switch.

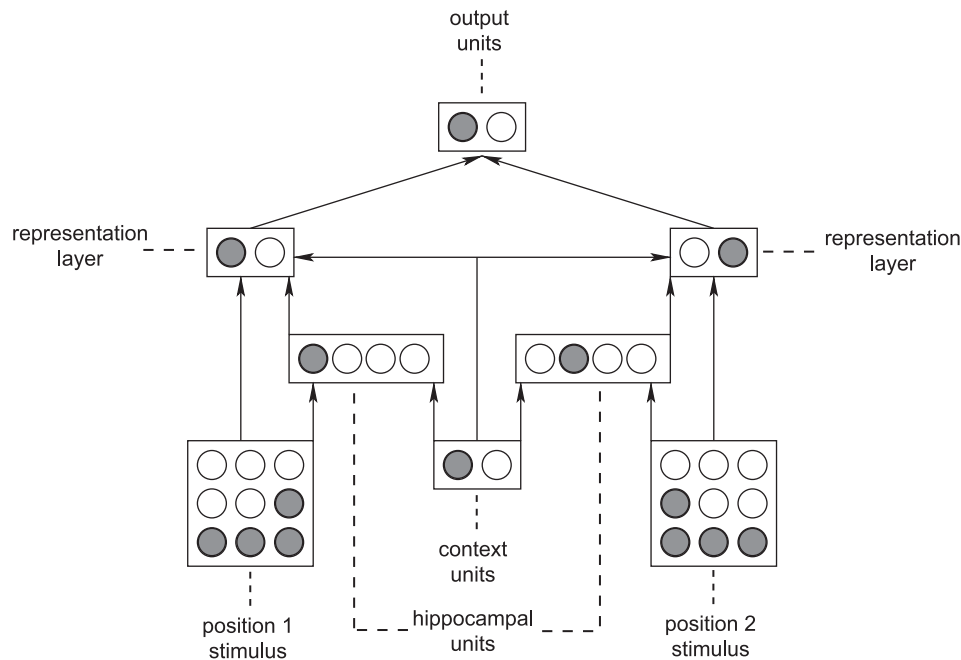
The 'baseline' condition in Experiment 4 was given to 16 children. The task required the child to select which of the three alternative actions could possibly elicit a certain effect. The results (Table 7) showed that children ordinarily choose an appropriate action over an inappropriate alternative. The 'test' condition, conducted with a separate group of 16 children, required appropriate actions on the part of the child to cause the effect to cease, after observing that the effect could be produced by the inappropriate action, and not by the appropriate action. For example, in the physical condition the children tended to say something like 'please stop' rather than remove the magnetic buttons.

Experiment 5 involved the same test with the same materials, applied to another group of 16 children, but this time after prior exposure to situations analogous to

**Table 7** Domain-appropriate and inappropriate responses in different conditions

Source	Response	Physical effects			Psychological effects		
		Baseline	Test	Transfer	Baseline	Test	Transfer
Schulz and Gopnik (2004)	Appropriate	1.00	0.00	0.75	1.00	0.00	0.69
	Inappropriate	0.00	0.75	0.25	0.00	0.81	0.31
	Other	0.00	0.25	0.00	0.00	0.19	0.00
Model	Appropriate	0.99	0.08	0.78	0.99	0.08	0.78
	Inappropriate	0.01	0.92	0.22	0.01	0.92	0.22

*Note:* Proportion of children choosing each type of response for Sobel *et al.* (2004), Luce ratio of activation of contextually appropriate output unit for appropriate and inappropriate inputs for the simulation.



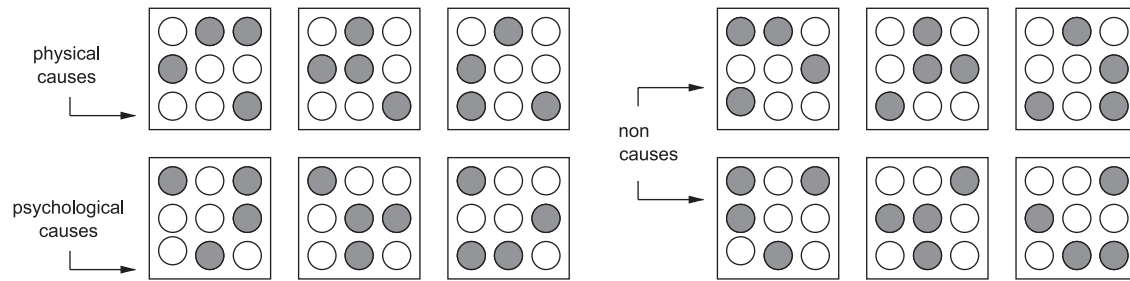
**Figure 4** The expanded structure used in simulation 3. Context and output layers each have a unit denoting psychological or physical context or observed effect. Arrows denote fully interconnected weight matrices.

the psychological and physical test situations, in which once again appropriate actions had no effect but inappropriate actions did have an effect. This condition is called the ‘transfer’ condition since the interest is in the extent to which prior exposure to the analogous situation would affect children’s performance in the test situation. The analogous pre-exposure situation for the physical task involved causing a box with a lucite top to light up using either of two rocker switches or by talking to the box. The analogous pre-exposure situation for the psychological task involved causing a person to ‘act goofy’ using either of two comical animal puppets or a switch. Prior to the transfer test, a test with these materials was conducted as in Experiment 4, demonstrating that here too, many children’s expectations were overridden (in the psychological condition, responses proportions were 0.81 inappropriate, 0.00 inappropriate and 0.19 ‘other’; in the physical condition, the proportions were 0.62 inappropriate, 0.19 appropriate, and 0.19 other). The results of the subsequent transfer test (Table 7), when compared to the baseline condition of Experiment 4, revealed that the recent relevant experience with similar objects in a similar context caused a slight tendency towards selecting domain-inappropriate actions: In Experiment 4’s baseline condition, children were completely consistent in choosing only the appropriate causes, whereas in the transfer test from Experiment 5,

about a quarter of the responses were domain-inappropriate actions.

#### Network structure

An expanded structure (Figure 4) incorporated an additional pool of input units to represent the domain or task context, and used two output units, one for each of the two effects. The premise is that a variety of objects have been experienced in various different contexts. In each context (Sobel *et al.*, 2004), actions of one type tend to produce a specific type of effect, while other types of actions will not produce the effect. Context is represented by activation of one of two context units; they serve here to indicate the type of effect desired (e.g. causing a machine to start or stop, causing someone to laugh or stop laughing). These play a role similar to task demand units in other networks (Cohen, Dunbar & McClelland, 1990), in that they determine what response is appropriate in a given situation. We call these contexts the psy-context and the phy-context, reminiscent of psychological and physical contexts of Schulz and Gopnik’s psychological and physical domains, but of course they are both completely abstract in the simulation. Of course this drastically oversimplifies the representation of context, which we view as multifaceted and distributed. The use of localist units for the two contexts here is consistent



**Figure 5** The stimuli used in cortical training of the network. Psychological and physical stimuli cause activation of their respective outputs when corresponding context units are active.

with the idea that they are quite different, and thus their representations are essentially non-overlapping.

#### Stimuli, training and testing

In cortical training, two types of stimuli were used (Figure 5) called phy-causes and psy-causes. Phy-causes are paired with the phy-effect in the phy-context, and psy-causes are paired with the psy-effect in the psy-context; when an input occurs in the wrong context, there is no effect. Phy-causes all had three units in common. Psy-causes had three *different* units in common. Each pattern also had one of the remaining three units on, such that each occurred once with a psy-cause and once with a phy-cause, and thus were indifferent with respect to the causal context. For testing and hippocampal training, the patterns used were drawn from those used in cortical pre-training. We believe the effects would generalize to similar but not pre-trained stimuli.

In the simulation of Experiment 4, the baseline condition assessed the tendency of contextually appropriate and inappropriate causes to activate the output based solely on cortical training. The model does not choose an action *per se*; rather, the network simply estimates whether a particular action (that is, placing a particular pattern on its input units) in a particular context will produce the appropriate effect. To relate these estimates to children's choices, we consider the possibility that children mentally simulate the effects of each of the available alternatives and choose among them with a probability proportional to the extent that each leads to the expectation that it will produce the outcome. To model this the output activations produced by the network for the alternative inputs are divided by the sum of the activations across both alternatives, normalizing the activations so that they sum to 1, following the Luce choice rule.

The exposure phase of the experiment was simulated in a manner similar to previous experiments, using weights from hippocampal units to the representation layer to

learn about the causal consequences of actions in the experimental situation, with all cortical connection weights frozen. There were two separate runs, one involving a psychological effect in a psychological context, and one involving a physical effect in a physical context. Each run was initialized with the same cortical weights used for baseline testing. In each run, the model was exposed to two events. In one event, one of the context units was activated, and an input previously appropriate to that context was presented without the corresponding effect. In the other, the same context unit was activated, and a domain-inappropriate input was presented with the effect appropriate for the context (and so not appropriate for the input). The domain-appropriate and inappropriate patterns were taken from the cortical training set. As in previous simulations, a hippocampal unit was assigned to each pattern used in this exposure phase, and weights from this unit to the representation layer were adjusted to produce a pattern on the representation layer corresponding to the given output. Following the exposure phase the test was administered, assessing the extent to which the patterns used during the exposure phase, when activated together with the corresponding hippocampal unit, now led to activation of the relevant effect.

The transfer test to simulate the corresponding condition of Experiment 5 was done after hippocampal training using another domain-appropriate and another domain-inappropriate pattern from the cortical training set. This meant that the patterns used for the transfer test (which were the same as those used in the baseline and test conditions described above) overlapped in  $\frac{3}{4}$  of their active units with the appropriate and inappropriate patterns previously seen by the network. The hippocampal unit that was activated with each of these patterns during hippocampal training was partially activated (0.5) during the transfer test, capturing the idea that the test pattern partially activated a memory for the related event from the exposure phase of the transfer experiment. The effect of this was to partially reinstate the learned hippocampal representation, thereby competing

with the cortically based tendency to produce activations consistent with pre-experimental experience. The extent of this competition can be regulated by the choice of the degree to which the prior representation is activated in the context; the value used in the reported simulation is a free parameter, whose value was chosen to approximate children's behavior reasonably closely while also being informative about the behavior of the simulation.

## Results

The empirical findings are compared with the simulation results in Table 7. Both the model and the children strongly prefer the contextually appropriate object in the baseline condition. Clearly the model, like children, has acquired context-specific knowledge based on past experience, and can use this knowledge to choose an appropriate action in each of the two contexts.

The results of the 'test' condition show that both the model and children are able to reverse their expectations based on the unexpected consequences observed during the exposure phase. Occasionally, children made responses classified as 'other': examples given of responses in this category are no response or saying 'I don't know'. We did not provide such an option for the model. While none of the children in the test phase of Experiment 4 chose an appropriate action, 9% of responses in the analogous test phase of Experiment 5 were appropriate (data from that condition are not presented in the table). Given the small numbers of responses involved (16 per cell for the experimental data) and the differences between the experiments and the simulation in the range of options available, the simulation and experiment seem to be producing comparable outcomes for the test condition.

The results of the generalization test show that, if the hippocampal memory of a related event is partially activated during testing, it can influence performance in the model, in a way that mirrors the effect of prior exposure to an analogous event in the children. It may be somewhat surprising that, in the model, half activation of the hippocampal unit produces a considerably less than 50% tendency to override previously established contextually appropriate expectations. This attests to the strength of these prior expectations in the model, as well as to the non-linearity in the model's operation. We do not wish to make much of the exact tendency shown by either the model or the children to generalize; surely this depends on details of the degree of similarity between the exposure and test situations that we could only speculate about and have no way of independently specifying for either the model or the experiment. Perhaps it is noteworthy, however, that the young children are not much more

likely than the model to override their pre-experimental knowledge arising from prior experience with contextually appropriate actions. Given that the situations used for the test and generalization conditions seem to have been designed with quite salient similarities, it seems that children, like the model, have relatively well-entrenched expectations that are not easily overridden by a single contrary experience in a similar situation.

## General discussion

Over three simulations, we have shown how a simple connectionist model based on general principles of learning and inference can address much of the evidence on which Gopnik *et al.* have based their argument that children possess a specialized module for causal inference. The model can reason from a few experiences and make indirect inferences; it can use prior information about relative rates of causal outcomes to modulate its probability estimates; and it can make use of prior beliefs about context-specific cause and effect relationships, but override them in specific cases when presented with contradictory observations.

We do not, of course, wish to suggest that this rules out the possibility that children do have a specialized causal reasoning module. A demonstration such as this can only indicate that an appeal to such a module may not be necessary. But before even this conclusion can be accepted, we must address several questions: Have we somehow snuck in a reliance on the features of a specifically causal formalism? Have we really addressed the essence of the causal inferences children are capable of making? Have we left out important aspects of children's use of causal maps that are inherently beyond the scope of connectionist approaches? We do not think so, but the questions deserve consideration.

*Does the model unintentionally depend on principles specific to causal inference?*

One might argue that the networks used in our simulation essentially presuppose a cause-effect relationship by having connections running forward from input units representing potential causes to output units representing potential effects. While our architecture does have this structure, it is the most generic form of architecture used in connectionist modeling, and is by no means specialized for causal inference. It is true that using the inputs for potential causes and the outputs for potential effects presupposes a specific direction of prediction, but we see this choice as reflecting a characteristic of the temporal structure of events that are conventionally

treated as involving causal relationships, including those situations we have considered in the experiments we have modeled. In these situations, a block is placed on a box and at the instant of contact the box begins to flash, or a noisy machine starts up immediately upon the occasion of someone speaking to it, and the network's input-to-output architecture simply mirrors this antecedent-to-consequent relationship. We suggest that even this architectural distinction is not necessary to learn appropriate predictions from antecedent to consequent events. Such predictions can also be learned in recurrent networks (Elman, 1990; Williams & Zipser, 1989) in which a representation of the state of the world based on a graded (and experience-dependent) representation of the events of the present and recent past is used to predict the state of the world at the next moment. In such networks, there is no architectural distinction between antecedent and consequent; the target of prediction at each time becomes part of the input for predicting what will be the target for prediction at the next point in time. We have argued elsewhere that such networks are well suited to addressing learning of causal relationships (Rogers & McClelland, 2004), but once again they are not specialized to this purpose. Indeed, one version of such networks was used by Elman (1990) to address the learnability of structure in a very different domain – the domain of syntax – another area in which it has been claimed that humans rely on a highly specialized module. The very same architecture is relevant equally to causal reasoning, syntax, and any context in which there is structure in temporal sequences.

There are other features of our network that play a role in its success, and here again it is relevant to consider whether these features bring in principles of a specifically causal nature. One of these is the use of weight sharing to allow the two input arrays in our network to be treated as equivalent for the purposes both of processing and learning. This architectural feature certainly contributes to the network's performance, but once again we do not see it as inherently related to causal as opposed to non-causal inference. Indeed, this idea of sharing weights across positions was first introduced by Rumelhart *et al.* (1986a) to allow position-independent object identification and generalization of learning about the relationship between appearance and identity across position, and methods for learning position-invariant object representations have been proposed (Foldiak, 1991; O'Reilly & Johnson, 1994), which could be used in causal reasoning tasks and other situations.

It may be worth pointing out that position can be highly relevant in causal situations. It is not that a blicket can be placed just anywhere, but that it can be placed anywhere on the surface of the blicket detector that is relevant in the present context. No doubt our

model has presupposed this. But we would argue that such a presupposition is completely distinct from the abstract causal Bayes net formalism that is attributed by Gopnik *et al.* (2004) to children. The causal Bayes network formalism is completely general with respect to particular types of causes including physical causes or causes of any other kind, and causes vary from case to case in their position dependence; compare the differences in position required for a key to open a lock, a remote control to operate a television, and for a magnet to align iron filings. So for each of these cases, an appeal must be made to a principle outside of the domain of causal reasoning as Gopnik *et al.* (2004) define it to specify the appropriate spatial relationship. As ever, we are open to the possibility that the spatial relationship required for the activation of the blicket detector (object X is on-top-of Y) might well be a learned representational category, based on regularities experienced in the environment. Among these, for example, is the regularity that placing an object on a flat surface has the usual consequence that it stays in position regardless of exactly where on the surface it is placed. (The question of whether this regularity is a causal, spatial, or physical regularity is, in our view, neither answerable nor necessary.)

In summary, while we acknowledge that there are properties of our model that are conducive to its successes, these are not properties one would associate specifically with causal reasoning. Instead, some of the properties are of general relevance in non-causal as well as causal reasoning situations, while others do not apply consistently across all causal reasoning situations, but are specific to particular sub-cases.

*But does the model capture the central properties of causal reasoning?*

An alternative to the possibility that our model incorporates an unintended appeal to principles specific to causal inference is the idea that it is not actually exhibiting causal reasoning at all. In this context it might be noted that our method for combining the outputs of the two positional slots in the network actually fails to do justice to the actual causal logic of the blicket/blicket detector mechanism. According to this logic, if any one of  $n$  objects placed on the detector is a blicket, it will activate; otherwise it will not: this is the logical OR relationship. Our cortical network uses graded summation of activation based on experiences with blickets and non-blickets. Based on the cortical pre-training in which the OR relationship holds, it comes to implement an approximation to this relationship. However, there are signs that it does not function as a perfect 'OR' network in all cases. For one thing, in simulation 2, it over-predicts

that an object participating in the AB+ event will be a blicket, relative to the estimate derived from the simple Bayesian analysis presented in the discussion of that simulation.

With the above in mind, one might be tempted to ask whether the successes of the model elsewhere are all an illusion. It might be suggested, then, that the model is getting by with a general-purpose form of inference that doesn't really correspond to the real essence of causal inference. We would argue against this suggestion. First, the particular form of the causal relation between an event and its consequent lies outside the scope of the causal Bayes net formalism; the logical OR relation among potential causes is only one possibility for the specific type of causal relationship. Second, the authors of Sobel *et al.* (2004) were satisfied that 4-year-old children's responses corresponded qualitatively to their expectations for their sensitivity both to the causal Bayes net schema and to the specific expectations one can derive from the logical OR relation using Bayesian inference, and the results of the simulation accord at least as well with one instantiation of such a Bayesian calculation (the one we offer in the discussion of simulation 2) as 4-year-old children's responses do. Third, the model we have offered in this paper might be enhanced by the adoption of a more explicit implementation of the logical OR relation for inference in this case, but such an enhancement would not be relevant to all causal inference situations (some involve graded accumulation of influences, as are quite naturally captured in general-purpose connectionist networks), nor would it be restricted to them as it would enhance equivalently the mechanism proposed by Rumelhart *et al.* (1986a) for position-invariant object recognition. We conclude that the network model, general-purpose as we claim it to be, actually provides an adequate match to the constraints specifically applicable to the causal Bayes net formalism.

A second complaint that might be leveled against the model is that we haven't dealt with the important distinction, within the Bayes net framework, of interventions vs. observations. Interventions are external inputs that directly manipulate the state of a system independent from those factors that operate within the system itself. It has been noted that interventions license different inferences than observations and that humans are sensitive to the difference.

While we do not wish to claim that we can address all aspects of situations researchers interested in causality treat as interventions, some of these situations do appear to fall within the purview of our approach. Consider the situations contrasted by Kushnir, Gopnik, Schulz and Danks (2003) in their consideration of this issue. In their experiments, a child sees two sticks protruding from a

box. In one condition, the child witnesses events in which both sticks move up and down together, and other events in which one stick moves and the other does not. In the other condition, there is what is called an intervention: the same events occur, but this time the independent movement events are produced by the experimenter reaching in and conspicuously grasping and moving one of the sticks (the other stick does not move in these cases). The responses to these situations are quite different: In the former case, people think the sticks are controlled by two independent (invisible) causes (which happen to occur together some of the time). In the latter case, people typically chose the interpretation that the sticks are controlled by a single invisible cause on the trials where the sticks move together. A further condition shows that it is not simply that the intervention provides a cue that can be associated with the independent movement, since if the experimenter simply points to each stick when it moves independently (rather than grasping, pulling and pushing it), two independent causes are still inferred. Apparently, if subjects can attribute the independent movement to a plausible external cause (such as grabbing and pulling/pushing the object), they will, leaving the cases where the objects move together without an external intervention attributable to a single common hidden cause. Otherwise, the cases of independent movement lead to the conclusion that there are two independent hidden causes even if they sometimes occur together.

We are not yet in a position to address the inference of unobserved causes. We believe our model can be extended to address such inferences; this would be accomplished by extending the idea of back-propagation to representation a step further, using it to construct a representation of something for which there is no direct evidence but whose existence can be inferred from what is observed. What we can do, however, is address a slightly simpler case – one which, we suggest, captures the essential feature of the intervention experiments. We suggest that an intervention is an instance of a situation in which a known cause for the state of a particular variable is apparent. Among other things, the presence of this known cause prevents the attribution of the cause of the variable's state to some other possible source. Consider a network that has participated in the cortical training phase of simulation 1, so that it 'knows' that the presence of objects with certain properties are associated with activation of its output unit but objects with other properties are not. Now consider what happens when an ambiguous object (we used the n1 object from simulation 1) is placed on the blicket detector, and the detector activates, but under two different conditions: (1) there is no other object on the detector; (2) there is another object on the detector (i.e. in the network's second input

slot) which is one of the objects that, in the network's experience, have previously been associated with detector activation. What happens is that, in the first case, back-propagation to activation assigns causal power to the ambiguous object, so that, when presented later on its own, it strongly activates the output (activation = .990). In the second case, the presence of the other object that already has the power to activate the output unit prevents this from happening. The network hardly alters its state of knowledge about the ambiguous block at all (activation = .353 after the experience, compared with a baseline value of .347 after cortical pre-training alone). What's happened here is that the network already has an object (the effective block) to which it can attribute the causal outcome; there is no error at the output units; and so there is no basis for adjusting the causal power it attributes to the ambiguous block. Finally, we can also consider the third scenario, analogous to the pointing condition of the hidden cause study described above, in which there is a second object, but it is one similar to the objects that, in the network's experience, have previously not caused the detector to activate. In this case, the network also strongly attributes causal power to the ambiguous object. After this experience, when this block is presented on its own, it activates the output almost to the same extent as in the case where the ambiguous block occurred on its own (activation = .978). Clearly, the network is able to attribute causation to a known cause when one is available, but attributes it to an ambiguous input when that occurs either by itself or with some other object that is not a plausible alternative cause, based on the network's prior experience.

We suggest that this demonstration shows that the network exhibits behaviors sufficient to describe it as understanding something about the causal structure of the world, and also shows that it can make appropriate use of such information to guide its causal attributions to ambiguous objects. We leave it to future research to investigate whether, as we suspect, the same kind of ability would be exhibited by a successor to this network that infers the existence of hidden causes using back-propagation to activation.

*Are connectionist models associative models, and does the model we present here stray beyond the bounds of these approaches?*

Gopnik *et al.* (2004) repeatedly contrast their approach with associative models, as do Sobel *et al.* (2004), but neither paper provides much consideration of connectionist models. However, connectionist models are sometimes lumped together with associationist models by others, and so, without attributing such a tendency to

Gopnik *et al.* (2004) or Sobel *et al.* (2004), we suspect it may be worth considering the relation.

Connectionist models are associative in that, like classical associative models, they learn from experiences involving co-occurrence or temporal succession and generalize on the basis of similarity. Crucially, though, as has been stressed elsewhere (Rogers & McClelland, 2004, for example), similarity in connectionist models is not inherent in the appearance of objects but depends on similarity relations among internal representations shaped by prior experience, including among other things experience with the consequences of their participation in events. The ability to form such internal representations is a consequence of the availability of multi-layer learning algorithms such as back-propagation. These algorithms were not available to classical associationists, and this is perhaps the most important way in which connectionist models expand on such approaches.

A difference between our model here and most associative and connectionist models is our use of rapid interleaving of relevant events to find representations for the different participating objects that effectively explain the outcomes of both events. Some readers may object that this stretches associationist and/or connectionist approaches beyond their natural boundaries. While we are not sure where to put the bounds on associationism, there is substantial precedent for appealing to mental iteration and interleaving in the connectionist literature. Connectionist models address the micro-structure of cognition, and work within this framework has long embraced the iterative use of networks at the macro-structural level, i.e. in temporally extended acts of cognition. For example, Rumelhart, Smolensky, McClelland and Hinton (1986b) proposed mental simulation as a natural way of extending the framework beyond mere reaction to the vicissitudes of external stimulation. It was suggested there, and we suggest again now, that such mental simulation is the connectionist (and might more broadly be seen as the associationist) implementation of thinking. Rumelhart used mental simulation to select appropriate next moves in a competitive game-playing situation. A use of interleaving similar to ours – to find a representation accounting simultaneously for different pieces of information – was previously used by Rogers and McClelland (2004). Also, interleaving was proposed by McClelland *et al.* (1995) to avoid the problem of catastrophic interference when integrating new information initially stored in the hippocampus into the cortical learning system. Thus, there is certainly a precedent for mental iteration and interleaving in the connectionist literature, where it has played an important role in several very distinct contexts.

Beyond this question of boundaries, it might still be objected that the specific instantiation of mental simulation

that we have employed – interleaved replay until equilibrium is reached – is rather far-fetched in terms of the processing demands it places on mental processes. Taken literally, we would have to swap events in and out repeatedly over many iterations. We share this concern – we do not imagine that a child or even an adult would literally do this kind of extensive interleaving in the context of experiments like those we have considered here (although we suspect this kind of thing may occur in the mind of a scientist mulling over a knotty scientific question over an extended time period). In this connection, however, it is worthwhile to note that full interleaving may not really be necessary, at least not in all cases. To illustrate this point, we returned to the retrospective disambiguation tasks considered in simulations 1 and 2, and investigated how the network used in these simulations would perform if the hippocampal training regime involved a single step of retrospective disambiguation. Specifically, we considered the possibility that a participant in the retrospective screening-off or backward blocking task is able to withhold updating on initial presentation of an ambiguous AB+ event; then assigns a hippocampally maintained representation to A based on the subsequent A+ or A– event; then holds this in mind and brings the AB+ event back to mind, updating hippocampal weights for both A and B.<sup>4</sup> After just this one step of retrospection, the output activation produced by the B pattern is .93 in the screening-off task, and .35 in backward blocking (very close to the baseline), approximating the good performance of 4-year-old children in the (Sobel *et al.*, 2004) experiments. Of course our proposal amounts to allowing the child to turn backward-into forward-disambiguation tasks, and considerable executive control might be required for this. But, once again, the procedure proposed would be relevant in non-causal as well as causal reasoning situations; essentially, we are simply breaking the dependence of the learning process on the literal sequence in which relevant information is presented to the learning system.

<sup>4</sup> For this we used the simulation 1 network as above, with these changes: We turned off momentum since this carries weight adjustments across phases; to compensate we increased the learning rate to .2. One hundred iterations were carried out for the A presentation and for the subsequent AB presentation. Note that these iterations do not involve interleaving, simply an iteration of forward and backward passes through the network, consistent with bi-directional propagation of activation while a fixed input and target are maintained. Note that the use of a fixed number of iterations is necessary: The network is effectively ergodic, and if the simulation is allowed to run all the way to equilibrium, in the presence of weight decay, the effects of the A presentation are washed out. This would not be necessary, however, if updating of the representation of A was prevented during the mental revisiting of the AB+ situation.

It should be noted that we do not propose that interleaving or retrospective mental representation of a prior experience occurs all the time. Rather, we suggest that it is likely to be invoked only in certain situations. For example, if exposures to events involving particular objects are well spaced out in time, and no specific demand is placed on a person experiencing these events to reason from them to a conclusion about the properties of the participating objects, prior events may not be brought back to mind and revisited with the consequences of the subsequent experiences for the representation of a subset of the objects used to help disambiguate earlier situations. But under the conditions of the blinket experiments, the exposure phase events are presented close to each other in time; the test phase follows immediately, and the child is probed for an inference about both of the objects that were used in the exposure phase. These conditions may invite explicit retrospective disambiguation. Such retrospective disambiguation is likely also to occur in situations where a current event prompts an episodic recollection of a previous related event.

In summary, our simulations suggest that extensions of classical associative learning models might provide a sufficient mechanism to explain many instances of human causal reasoning. The required extensions include the use of a sophisticated learning algorithm such as back-propagation to allow expectations to be based on internal representations structured by experience, as well as mental reconsideration of prior inputs after representations have been adjusted based on subsequent information. Whether such models should still be called associationist or not, they certainly have clear links to associationist approaches.

## Conclusion

A common tendency in cognitive and psychological research is to invoke something special – a cognitive module perhaps or a specific type of core knowledge – to address each different type or domain of cognitive competence we see reflected in people's judgments and behavior. Among others, modules for syntax, vision, numerosity, theory of mind, and more recently causal reasoning have been invoked to explain the particular remarkable cognitive abilities of adults and even young children. Perhaps it is tempting to invoke such modules, since each domain apparently operates according to its own principles. Our simulations suggest, however, that there may be reasons to continue to consider the alternative idea that apparently domain-specific capabilities arise from mechanisms embodying a shared set of principles operating on different data. For this approach to work, the mechanisms

that embody these principles must necessarily be more powerful, and more structured, than historic approaches based on classical associative learning algorithms. We have relied on more powerful multi-layer learning algorithms, complementary learning systems, and retrospective reconsideration of prior experiences in our simulations. We think it likely that future research will bring out the need for yet further extensions, as the effort to discover the domain-general principles of cognition is pursued.

## Acknowledgements

Supported by MH64445 to JLM. The authors would like to thank David Danks, Alison Gopnik, Clark Glymour, Gary Lupyan, Punitha Manavalan and Gautam Vallabha for helpful and relevant discussions and/or pointers to relevant work. We also wish to thank Sindhu John and Punitha Manavalan for implementing software extensions and help with the simulations.

## References

- Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.
- Bülthoff, H.H., & Yuille, A.L. (1996). A Bayesian framework for the integration of visual modules. In T. Inui & J.L. McClelland (Eds.), *Attention and performance XVI: Information integration in perception and communication* (pp. 49–70). Cambridge, MA: MIT Press.
- Carey, S. (1978). The child as word learner. In M. Halle, J. Bresnan, & G. Miller (Eds.), *Linguistic theory and psychological reality* (pp. 264–292). Cambridge, MA: MIT Press.
- Cohen, J.D., Dunbar, K., & McClelland, J.L. (1990). On the control of automatic processes: a parallel distributed processing model of the Stroop effect. *Psychological Review*, **97**, 332–361.
- Elman, J.L. (1990). Finding structure in time. *Cognitive Science*, **14**, 179–211.
- Foldiak, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, **3**, 194–200.
- Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: causal maps and Bayes nets. *Psychological Review*, **111**, 3–32.
- Gopnik, A., & Sobel, D. (2000). Detectingblickets: how young children use information about novel causal powers in categorization and induction. *Child Development*, **71** (5), 1205–1222.
- Gopnik, A., Sobel, D., Schulz, L., & Glymour, C. (2001). Causal learning mechanisms in very young children: two-, three- and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology*, **37**, 620–629.
- Gopnik, A., & Wellman, H. (1994). The theory theory. In L. Hirschfeld & S. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 257–293). New York: Cambridge University Press.
- Hinton, G.E., & McClelland, J.L. (1988). Learning representations by recirculation. In D.Z. Anderson (Ed.), *Neural information processing systems* (pp. 358–366). New York: American Institute of Physics.
- Krushke, J.K., & Blair, N.J. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin and Review*, **7**, 636–645.
- Kushnir, T., Gopnik, A., Schulz, L., & Danks, D. (2003). Inferring hidden causes. In *Proceedings of the 25th Annual Meeting of the Cognitive Science Society* (pp. 699–703). Mahwah, NJ: Erlbaum.
- McClelland, J.L. (1986). The programmable blackboard model of reading. In J.L. McClelland, D.E. Rumelhart, & PDP Research Group (Eds.), *Parallel distributed processing: Volume 2: Psychological and biological models* (pp. 122–169). Cambridge, MA: MIT Press.
- McClelland, J.L. (1996). Integration of information: Reflections on the theme of attention and performance. In T. Inui & J.L. McClelland (Eds.), *Attention and performance XVI: Information integration in perception and communication*. Cambridge, MA: MIT Press.
- McClelland, J., McNaughton, J., & O'Reilly, R. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, **102**, 419–457.
- McClelland, J., & Plaut, D. (1999). Does generalisation in infant learning implicate abstract algebra-like rules? *Trends in Cognitive Sciences*, **3** (5), 166–168.
- McClelland, J., & Rumelhart, D. (1986). A distributed model of human learning and memory. In J.L. McClelland, D.E. Rumelhart, & PDP Research Group (Eds.), *Parallel distributed processing: Volume 2: Psychological and biological models* (pp. 170–215). Cambridge, MA: MIT Press.
- McNaughton, B.L., & Morris, R.G.M. (1987). Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends in Neurosciences*, **10**, 408–415.
- Miikkulainen, R., & Dyer, M. (1987). Building distributed representations without microfeatures. Technical report, Artificial Intelligence Laboratory, Computer Science Department, University of California, LA.
- Mozer, M.C. (1987). Early parallel processing in reading: a connectionist approach. In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading* (pp. 83–104). London: Erlbaum.
- O'Reilly, R.C. (1996). Biologically plausible error-driven learning using local activation differences: the generalized recirculation algorithm. *Neural Computation*, **8**, 895–938.
- O'Reilly, R.C., & Johnson, M. (1994). Object recognition and sensitive periods: a computational analysis of visual imprinting. *Neural Computation*, **6**, 357–389.
- O'Reilly, R.C., & McClelland, J.L. (1994). Hippocampal conjunctive encoding, storage, and recall: avoiding a tradeoff. *Hippocampus*, **4**, 661–682.
- Pearl, J. (2000). *Causality*. New York: Oxford University Press.
- Piaget, J. (1930). *The child's conception of physical causality*. New York: Harcourt, Brace.

- Rogers, T.T., & McClelland, J.L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.
- Rohde, D. (1999). Lens: the light, efficient network simulator. <http://tedlab.mit.edu/dr/Lens>.
- Rumelhart, D., Hinton, G., & Williams, R. (1986a). Learning internal representations by error propagation. In D. Rumelhart, J. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructures of cognition* (Vol. 1, chapter 8, pp. 318–362). Cambridge, MA: MIT Press/Bradford Books.
- Rumelhart, D.E., Smolensky, P., McClelland, J.L., & Hinton, G.E. (1986b). Schemata and sequential thought processes in PDP models. In J.L. McClelland, D.E. Rumelhart, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2, chapter 14, pp. 7–57). Cambridge, MA: MIT Press.
- Rumelhart, D., & Todd, P. (1993). Learning and connectionist representations. In D. Meyer & S. Kornblum (Eds.), *Attention and performance XIV* (chapter 1, pp. 3–30). Cambridge, MA: MIT Press/Bradford Books.
- Schmajuk, N.A., & Larrauri, J.A. (2006). Experimental challenges to theories of classical conditioning: application of an attentional model of storage and retrieval. *Journal of Experimental Psychology: Animal Behavior Processes*, **32**, 1–32.
- Shultz, T.A. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development*, **47**, 1–51.
- Schulz, L., & Gopnik, A. (2004). Causal learning across domains. *Developmental Psychology*, **40** (2), 162–176.
- Sobel, D., Tenenbaum, J., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science*, **28**, 303–333.
- Van Hamme, L.J., & Wasserman, E.A. (1994). Cue competition in causality judgments: the role of nonpresentation of compound stimulus elements. *Learning and Motivation*, **25**, 109–121.
- Williams, R.J., & Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, **1**, 270–280.