Causality and Imagination

Caren M. Walker & Alison Gopnik

University of California, Berkeley

Abstract

This review describes the relation between the imagination and causal cognition, particularly with relevance to recent developments in computational theories of human learning. According to the Bayesian model of human learning, our ability to imagine possible worlds and engage in counterfactual reasoning is closely tied to our ability to think causally. Indeed, the purpose and distinguishing feature of causal knowledge is that it allows one to generate counterfactual inferences. We begin with a brief description of the "probabilistic models" framework of causality, and review empirical work in that framework which shows that adults and children use causal knowledge to generate counterfactuals. We also outline a theoretical argument that suggests that the imagination is central to the process of causal understanding. We will then offer evidence that Bayesian learning implicates the imaginative process, and conclude with a discussion of how this computational method may be applied to the study of the imagination, more classically construed.

Keywords

Causal Reasoning, Bayes Nets, Imagination, Fictional Cognition, Counterfactuals

Introduction (h1)

Conventional wisdom suggests that knowledge and imagination, science and fantasy, are deeply different from one another – even opposites. However, new ideas about children's causal

reasoning suggests that exactly the same abilities that allow children to learn so much about the world, reason so powerfully about it, and act to change it, also allow them to imagine alternative worlds that may never exist at all. From this perspective, knowledge about the causal structure of the world is what allows for imagination, and what makes creativity possible. It is because we know something about how events are causally related that we are able to imagine altering those relationships and creating new ones.

A large portion of our psychological lives is spent engaging in counterfactual thought, planning and anticipating our future states, and considering near and far alternatives to the actual state of the world. While the imagination has long been assumed to generate counterfactuals (see Harris, German, & Mills, 1996; Beck & Riggs, this volume), little research has explored *how* human minds, even the very youngest human minds, manage to produce these counterfactuals, how we know which possibilities will be the most likely to occur, and why imagining new possibilities is important.

In this chapter, we propose that part of the answer is that our ability to imagine possible worlds is closely tied to our ability to think causally. Recent work emphasizes the close two-way relationship between causal and counterfactual thought. Indeed, the purpose of causal knowledge, and the feature that distinguishes it from other kinds of knowledge, is that it allows you to generate counterfactual inferences. In particular, causal knowledge is useful, both ontogenetically and evolutionarily, because it allows for a special kind of counterfactual called an intervention. Once you know how one thing is causally connected to another, this knowledge allows you to deliberately do things that will change the world in a particular way. Intervening deliberately on the world isn't the same as just predicting what will happen next. When we intervene, we envision a possible future we would like to bring about, and our action actually

changes the world. Having a causal theory makes it possible to consider alternative solutions to a problem, and their consequences, before actually implementing them, and it facilitates a much wider and more effective range of interventions. This kind of sophisticated and insightful planning, then, involves a particularly powerful kind of imaginative capacity, and is tied to causal knowledge.

Counterfactuals also play a role in learning causal knowledge itself. A Bayesian view of learning (e.g., Griffiths & Tenenbaum, 2005; Griffiths, et al., 2010) suggests that children learn by generating possible patterns of evidence from alternative models in order to assess the fit between the outcome of these alternatives and the actual evidence. Just as causally sophisticated planning involves a kind of imagination, so does the very process of learning on a Bayesian view. In both cases, children must generate patterns of evidence from a premise that is not currently held to be true.

This review focuses on relatively recent developments in the field of computational theories of human learning that have arisen over the past decade (e.g., Pearl, 2000; Spirtes, Glymour, & Scheines, 1993; Tenenbaum & Griffiths, 2003; Woodward, 2003), and much of the evidence that will be presented is research that links observation, intervention, and counterfactual reasoning as sharing a common foundation in causality. We will therefore begin with a brief description of notions of causality in the "probabilistic models" framework, and outline a theoretical argument that suggests that the imagination is central to causal understanding. We will then offer some evidence that suggests that human children are indeed rational Bayesian learners of causal models and discuss how this learning also implicates imaginative processes. We conclude with a discussion of how this work may be applied to the study of the imagination, more classically construed, such as the study of pretend play and imaginary companions. We

will describe some research that is currently underway, and offer suggestions for future work to more directly examine the development of imagination from a Bayesian perspective.

A Bayesian Picture of Causality (h1)

Causal learning is a notorious example of the gap that exists between our experience of events and the truth. The philosopher David Hume (1748) originally articulated this difficulty: all we see are contingencies between events – one event follows another. How do we ever know that one event actually caused the other? To make matters more difficult, causal relations are rarely limited to just two events. Instead, dozens of different events are related in complex ways, and deterministic causal events are very rare – often, a cause will make an effect more likely, but not absolutely certain (e.g., smoking causes lung cancer, but not always, and whether a particular smoker actually gets cancer depends on a complex web of other factors).

Historically, psychologists believed that young children were precausal (Piaget, 1929). However, over the past twenty-five years, there has been a growing body of research that suggests that by the age of five, children understand a great deal about the complexities of the causal world, including the principles of everyday physics (e.g., Bullock, Gelman & Baillargeon, 1982; Spelke, Breinlinger, Macomber, & Jacobson, 1992), biology (e.g., Gelman & Wellman, 1991; Inagaki & Hatano, 2006), and psychology (e.g., Gopnik & Wellman, 1994; Perner, 1991). By 2-years of age, children begin to make causal predictions and provide causal explanations for physical phenomena in the world (e.g., Legare, Gelman, & Wellman, 2010; Hickling & Wellman, 2001), for the actions of others (e.g., Wellman & Liu, 2007), and even for imaginary or counterfactual scenarios (e.g., Harris, German, & Mills, 1996; Sobel & Gopnik, 2003). Further, because this causal knowledge has been shown to change in the face of new evidence (e.g., Gopnik, Glymour, Sobel, Schulz, Kushnir, & Danks, 2004; Slaughter, Jaakkola, & Carey,

1999; Schulz, Bonowitz, & Griffiths, 2006), it seems that causal knowledge is learned, undergoing change over the course of development (e.g., Gopnik & Meltzoff, 1997; Gopnik et al., 2004).

Developmental theory theorists' (e.g., Carey, 1985; Gopnik, 1988; Gopnik & Meltzoff, 1997; Wellman, 1990; Wellman & Gelman, 1998) have proposed that this causal knowledge is represented by a set of theories that are revised over the course of development in a process that is analogous to scientific theory-formation and revision. These theories support abstract causal reasoning and therefore enable the learner to make predictions, provide explanations, and even reason about counterfactuals in a variety of domains. Recently, that there has been major progress towards building a precise computational theory describing the representations and learning mechanisms that may account for theory change.

Much of this progress is a product of the current revolution in cognitive science concerning the rise of "probabilistic modeling" accounts of reasoning and learning (Chater, Tenenbaum & Yuille, 2006, Griffiths, et al., 2010; Pearl, 2001; Glymour, 2003). Many of the ideas about probability that underpin these models were first formulated by the philosopher and mathematician, Reverend Thomas Bayes, in the 18th century, and are now being successfully applied to a very broad set of problems in developmental psychology, including induction and inference in learning (e.g., Glymour, 2003; Gopnik & Schulz, 2007; Tenenbaum, Griffiths, & Kemp, 2006), language acquisition in infancy (e.g., Chater & Manning, 2006; Tenenbaum & Xu, 2000; Xu & Tenenbaum, 2007; Niyogi, 2002; Dowman, 2002; Regier & Gahl, 2004), and the development of social cognition (e.g., Goodman, Baker, Bonawitz, Mansinghka, Gopnik, Wellman, Schulz, & Tenenbaum, 2006; Baker, Saxe, & Tenenbaum, 2006), among others. The

application of probabilistic models has successfully described and predicted patters of behavioral data across a variety of experimental paradigms.

The idea that is foundation for all of this work is that learning is based upon the *assessed probabilities of possibilities*. According to a Bayesian account, nothing is ever certain; instead, we form rational inferences based upon the fact that some possibilities are more likely than others. As we accumulate more evidence about the underlying causal structure of the world, we systematically update the likelihood of all those possibilities. Therefore, a very small amount of evidence can effectively support one hypothesis over another. Similarly, if the evidence is strong enough, even the most unlikely possibility can turn out to be true, regardless of our previous experience or theories about the world. The process of learning therefore represents the movement towards more informed inferences that better approximate the truth in a broader range of scenarios.

This computational work in Bayesian inference has also begun to specifically examine the mechanisms that may underlie children's learning about the causal structure of the world. There has been particularly impressive progress in constructing representations of causal structure that are well suited to Bayesian learning. This work has begun to provide a solution for the problem of causal induction: how we derive rich, abstract representations from the sparse, concrete data that is available in our environment. More specifically, these accounts describe a mechanism that allows theory-like knowledge to be derived from data in our environment while also explaining how our prior knowledge constrains the inferences that we make, and the evidence that we choose to attend to. By actively uncovering the underlying causal structure of the world from evidence in the environment, causal learning may be conceptualized as the dynamic mechanism that underlies the process of theory change; the application of a theory to a

pattern of evidence is the process of assigning a particular causal representation to that evidence (Gopnik, 2000). Children converge on the truth following multiple iterations of major conceptual changes in which existing theories about causal structure are revised, and eventually abandoned and replaced over the course of development.

According to this account, children's brains construct a kind of unconscious causal map, or a picture of the way the world works. Many animals, from rats to human beings, construct "cognitive maps" of the spatial world, internal pictures of where things are in space (Tolman, 1948). Once spatial information is represented in this way, the learner is able to use that information much more flexibly and productively. A map is a very efficient device for constructing different cognitive blueprints, pictures of what will happen as you move yourself around through space, and this facilitates the consideration of many complex spatial possibilities before committing to any particular course of action. These spatial maps provide a coherent, non-egocentric, complex representation of the spatial relations among objects in the environment (O'Keefe & Nadel, 1979). As we move through the world, we are able to update this information to reflect newly learned input about the layout of the environment. Human beings also construct a different kind of map - a map of the complex causal relations among events that exist in the world (see Campbell, 1994; Gopnik, 2000). These causal maps, which may be unique to human cognition, share many of the advantages of spatial maps – allowing us to represent the relations among objects, independent of our own actions.

Causal Bayes Nets (h2)

Ideas about the role of causal maps in human learning emerged roughly two decades ago, when a group of philosophers led by Clark Glymour and a group of computer scientists led by Judea Pearl simultaneously began formulating a mathematical account of how theory change

might work in artificially intelligent systems. The mathematical descriptions that they produced were called "causal graphical models," more commonly referred to as "Bayes nets" (Pearl, 2000; Spirtes, Glymour, & Scheines, 1993, 2001). This work has transformed the field of artificial intelligence and inspired new ideas about causation in philosophy. Recently, developmental theory theorists, philosophers, and computer scientists have combined their efforts to describe how theory change could occur through the accumulation of knowledge that is represented in these Bayes nets, which collectively create causal maps of the known world (Gopnik, et al., 2004; Gopnik & Schulz, 2007).

Bayes nets represent causal relationships in directed acyclic graphs (see Figure 1). These causal models represent a normative mathematical formalism that provides a way of representing causal structure, as well as a set of tools for making accurate predictions and effective interventions on the environment to uncover the underlying causal structure of the world. Nodes in the graph represent the observable or hidden variables in a particular causal system, and the arrows represent the directed relationship that exists between these variables. These relationships can take a variety of different functional forms: deterministic or probabilistic, linear or non-linear, generative or inhibitory. The "parameterization" of the graph informs us about the nature of these functional relationships in more detail, including specific information about the relationships between nodes.

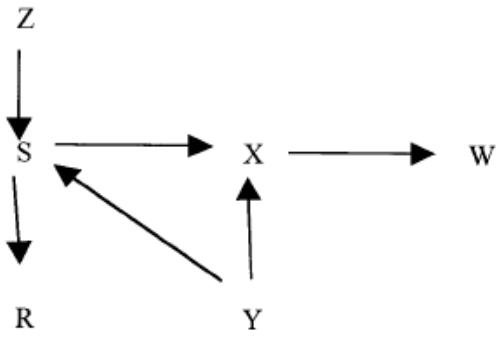

Figure 1. A causal graph (Reprinted with permission from Gopnik, et al., 2004)

The entire graph defines a joint probability distribution that exists over all the variables in the network – a distribution that describes the likelihood of the relationship between each of the variables. For example, the graph above can tell us something about the probability of a value of W given a value of R and S. The structure of these directed graphs therefore encodes probabilistic relations between variables that are updated based upon observed and inferred events that take place in the environment. Causal Bayes nets are therefore able to facilitate human reasoning about the potential effects of our causal actions, because knowledge about the underlying causal structure permits the learner to make a range of predictions about future events.

There are three basic causal structures that have been examined in the majority of research to date (see Figure 2): (1) common-cause models, in which a single cause X influences two effects Y and Z, (2) causal-chain models, in which an initial cause X affects an intermediate event Y that then influences the later effect Z, and (3) common-effect models, in which two causes X and Y independently influence a single effect Z. More complex models may be constructed via combinations of these three.

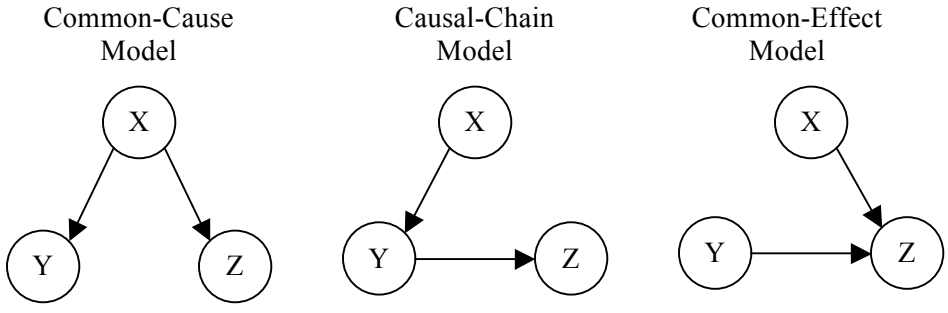

Figure 2. Three basic causal models (reprinted with permission from Hagmayer, Sloman, Lagnado, & Waldmann, 2007)

By specifying the probability distributions of the events within the graph, it is possible to make predictions about related events, with each distinct type of causal structure (common-cause, causal-chain, or common-effect) supporting a unique pattern of predictions. By encoding assumptions about dependence and independence among the represented variables in a directed graph, Bayes nets are able to provide the learner with a simplified representation of a particular causal domain

More recent work (Griffiths & Tenenbaum, 2007) has extended the basic idea of causal Bayes nets to Hierarchical Bayes Nets. These representations capture higher-order generalizations about specific Bayesian graphical models. To take an example from intuitive psychology, we may want to represent the higher-order fact that combinations of desires and beliefs cause actions. We could do this by including the constraint that all Bayes nets that involve mental state variables will have arrows that go from beliefs and desires to actions, without specifying what those particular causal relationships will be. Hierarchical Bayes Nets can therefore capture the idea of "higher-order" framework theories, or representations that are more abstract than specific theories, in a computational way (Wellman and Gelman, 1998).

For the last ten years, developmental cognitive scientists have explored the hypothesis that children represent causal relationships implicitly in the form of causal Bayes nets and learn new causal representations from observations of correlations and interventions. According to this hypothesis, as children develop, they actively fill in the probabilities associated with causal events. Learning the causal structure that is represented in Bayes nets requires an associated learning algorithm that includes *a priori* beliefs about what constitutes a plausible cause, and expectations about how a given causal structure leads to observed events in the world (e.g., Gopnik, et al., 2004; Gopnik & Tenenbaum, 2007; Schulz, Bonowitz, & Griffiths, 2007).

According to the Bayesian account, the learner optimizes their performance by constructing a hypothesis space H of all possible causal models. Given some data d, the learner computes a posterior probability distribution $P(h \mid d)$ representing a degree of belief that each hypothesis h corresponds to the actual causal structure of the world. These posteriors depend upon the prior probability $P(h)$ and the likelihood $P(d \mid h)$ that you would observe d if h were true. This computational approach is summarized in Bayes' rule:

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h'} P(d|h')P(h')}.$$

This formula specifies how posterior conditional probabilities of a particular hypothesis being true (given the data) are computed from the prior probability of the hypothesis multiplied by the likelihood of those data assuming the hypothesis is true. Because Bayesian learning uses structured priors and likelihoods that are drawn both from the learner's background or innate knowledge about causal structure, as well as observed or hypothetical data, variations on this simple algorithm provides a natural framework in which to consider how children modify their existing knowledge in the face of new evidence[1].

A typical Bayesian causal learning algorithm may proceed as follows. Take the current best hypothesis about the world (i.e., reality). Modify that hypothesis to produce an alternative hypothesis (or several). Generate the probability distribution for evidence that would result from that modified model, and do the same for the actual model. Then use a Bayesian inference procedure to compare the probability of the actual evidence under the previous hypothesis and the new modified hypothesis. If the posterior probability of the new hypothesis is greater, accept that hypothesis.

Combining this Bayesian learning algorithm with causal Bayes nets offers an extraordinarily powerful means for optimized learning in both artificially intelligent systems and human minds. If the learner has two different possible theories about the world – two possible causal maps – Bayesian inference may be used to select the more likely of these two possibilities, and use this inference to motivate rational action. Using the correct map facilitates accurate predictions. For example, if I think that smoking causes cancer, I can predict that preventing smoking will lower the probability of cancer. If it doesn't cause cancer – if the causal map is different – then preventing smoking won't have this effect, and I can use this information to generate a more accurate causal map of this relationship. If, however, the causal map successfully predicted the evidence that I observe, then the probability that that is the correct map will go up when making future inferences. Observing new evidence therefore makes one map more likely than another: if the likelihood of cancer goes down when people stop smoking, the likelihood that smoking causes cancer goes up. Causal maps therefore provide the learner with a method for making predictions about the structure of the world. By comparing those predictions with what is actually observed, the learner is able to systematically determine the likelihood that a particular causal map is actually true.

To summarize the Bayesian perspective on causal learning, children undergo a series of major conceptual changes in which early theories are abandoned and replaced over the course of development. This theory change is facilitated via updating the probabilistic independence and conditional independence relations among the variables that are represented in the causal Bayes net, and the resulting probability associated with the underlying causal model. These representations and associated learning algorithms allow the individual to both learn causal

structure from patterns of evidence and to predict patterns of evidence from their existing knowledge of causal structure.

Even more important however, using directed graphs to represent knowledge about causal structure allows the learner to *intervene* on a particular variable within a causal system. These interventions lead to predictable changes in the probabilistic dependencies over all of the variables that exist in the causal structure, thus allowing the learner to explore the contingencies that exist between nodes of the graph. Access to representations of causal structure therefore not only allows the learner to make wide-ranging predictions about future events, but also provides the means for intervening on the environment to bring about new events or imagine novel ways of arranging the world.

Causal Relationships Imply Counterfactuals (h1)

While the imagination may not fit with earlier notions of causality, the Bayesian account of causality suggests that imagination and the consideration of counterfactual possibilities are central to causal structure. A fundamental feature of Bayes nets is that they allow the learner to go beyond the way that the world actually is and engage with possibility – the way that the world could be. Philosophers (Lewis, 1986; Mackie, 1974) have long suggested that new causal relationships are learned by explicitly engaging with counterfactual alternatives. According to the interventionist account that is implicit in computational models like causal Bayes nets, causal relations may be understood in terms of a counterfactual claim: the proper interpretation of the claim *X causes Y* is that, all else being equal, *if you intervened to change X that would lead to a change in Y*. The causal arrows in a Bayes net are therefore defined in terms of possible interventions. These interventions need to be conceivable, though not necessarily feasible; a point that differentiates this account from earlier notions of causality based upon generative

transmission (e.g., Shultz, 1982). Because the causal relationships exist between individual variables, every existing causal relationship implies the existence of a related counterfactual.

In fact, this is the precisely the point that distinguishes causal relationships from simple correlations. While a correlation indicates that two events co-occur, a causal relation has the additional requirement of counterfactual dependence: in some other possible world, in which the cause had not initiated the underlying causal mechanism, the effect would not have occurred at all. By identifying the causal structure, causal Bayes nets support inferences about the effects of real and imagined interventions on the variables (Schulz, Kushnir, & Gopnik, 2007). Knowing the causal graph therefore allows you to predict the outcome of interventions, regardless of whether you have ever observed them being performed or even whether they *could* ever be performed. Causal relationships actively generate possible worlds, some of which are factual (they exist) while others are counterfactual (they do not exist). While we do not necessarily engage in conscious tracking of counterfactuals when we are reasoning causally, the activity of imagining brings these underlying counterfactuals to the surface (Sloman, 2005). Imagining new possible worlds may therefore be defined as the process by which implied counterfactuals become explicit.

The implications of this account were tested in a series of experiments to see whether children could use the patterns of dependence and independence among variables to infer causal structure, make novel predictions and perform appropriate counterfactual interventions (Gopnik, Sobel, Schulz, & Glymour, 2001; Schulz & Gopnik, 2004). In one such experiment, preschoolers were shown three flowers that were sometimes associated with a puppet sneezing. One flower (A) make the puppet sneeze 100% of the time, while the other flowers (B and C) only made the puppet sneeze when presented along with flower A (sneezing was unconditionally

dependent upon flower A; sneezing with flower B and C was independently conditional upon flower A). This structure allows the learner to make predictions about possible interventions. An intervention on flower A will change the probability of B and C generating the sneeze, but an intervention on flower B or C will not have this same effect. When children were asked to ensure that the puppet did not sneeze, children successfully used the pattern of conditional dependence and independence and removed flower A from the bunch. Note that they did this even though they had never seen the flower removed before nor observed the consequences of that action. To solve this problem they had to first infer the right causal model from the data, and then actively use that model to generate the imagined consequences of a possible action.

The "*do* operator" (h2)

The causal modeling framework provides a means for representing these interventions. Interventions alter the structure of the causal model by cutting the manipulated variable from its usual causes. To do so, interventions are treated as variables with certain special properties: they cannot be influenced by other causal factors in the graph, and they act to fix the value of the variable of interest. Pearl (2000) developed a mathematical means for representing an intervention via the inclusion of a "*do* operator." The *do* operator modifies a causal graph by disconnecting a variable from its parent causes, thereby nullifying the influence of these causes, while keeping the rest of the graph intact. Once the variable is cut from its causes, the intervention acts to set the value of the variable in a particular causal model: we use the *do* operator to set X to some value *x*, or $do(X = x)$. The effects of an intervention are then computed using the probability calculations on this "mutilated" graph in the same way that it would be computed on the original graph (see Pearl, 2000; Sprites, Glymour, & Scheines, 1993 for a

detailed description of the mathematics, or see Sloman, 2005 for an accessible overview). Pearl (2000) aptly refers to this intervention as performing "graph surgery" (see Figure 2).

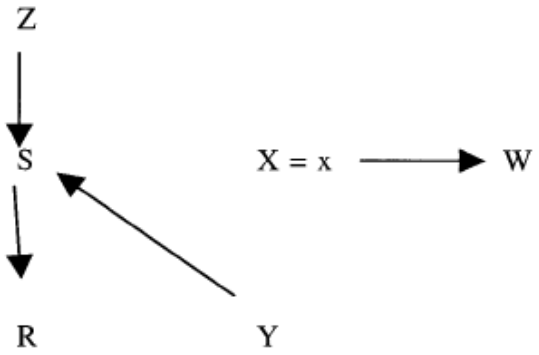

Figure 2. The "mutilated" version of the graph that appeared in Figure 1, following an intervention on X (reprinted with permission from Gopnik, et al., 2004)

Knowing that X is a direct cause of Y means that if the rest of the causal graph is held constant then intervening to change X to *x* should change Y to some value *y* (Pearl, 2000; Spirtes, Glymour, & Scheines, 1993; Woodward, 2003). There is substantial evidence that adults learn causal relationships more quickly and efficiently when they are able to perform these types of interventions, rather than relying on observation alone (Lagnado & Sloman, 2004; Sobel & Kushnir, 2006). These interventions enable the learner to differentiate among the possible causal structures that match the evidence that we observe in the world by creating a scenario in which the occurrence of an event is independent from its normal causes. While predictions regarding the outcome of observed phenomena are based upon the parameters outlined in the original causal graph, predictions regarding the outcome of interventions are based upon the parameters that appear in the mutilated causal graph. Therefore, a central feature of causal Bayes net learning algorithms is that they works in both directions: you can use evidence from interventions to infer the underlying causal structure *and* you can use knowledge of causal structure to predict the outcome that should result from an intervention. While observation and

intervention support different predictions and inferences, both are derived from the same basic process in the Bayesian model of causal learning.

While graphical surgery allows for the representation of actual interventions that take place in the physical world, this same process also provides a means for representing counterfactual interventions – imagined interventions that take place inside our heads. In counterfactuals the current state of the world is modeled and then intervened upon to construct the fictional scenario. Several studies have shown that adults can make accurate inferences about the effects of hypothetical interventions based upon information about causal structure (Sloman & Lagnado, 2005; Waldman & Hagmayer, 2005).

In one such study, Waldmann & Hagmayer (2005) assessed adult's ability to derive predictions for hypothetical interventions from fictitious causal models presented in a medical scenario. Adult participants were presented with one of two possible causal models: either a common-cause model or a causal-chain model. In the common-cause model, participants were informed that raising the level of hormone P in a chimpanzee causes the level of hormones S and X to increase. In the causal-chain model, participants were informed that raising the level of S in a chimpanzee causes P to increase, which then causes X to increase as a result of the increase in P. After this learning phase, participants were provided with observational data for 20 chimpanzees that illustrated the probabilistic relationships between the causes and effects.

All participants were able to successfully use this data to assess the parameters of each causal model in making predictions about hypothetical observations and interventions. More tellingly however, participants were highly sensitive to the different predictions that are generated for observations and interventions on each of these distinct causal structures. For example, when asked to imagine that the chimpanzees were injected with a substance that either

increased or lowered the level of S (an intervention on S), participants were able to differentiate their predictions between the two models. In the causal-chain model, increased levels of S would lead to a higher probability of increased levels of P and X, regardless of whether this increase in S was observed to naturally occur or due to the intervening effects of an injection. In the common-cause model however, participants demonstrated their understanding of the dissociation between instances when increased levels of S were observed to naturally occur and instances when increased levels of S were created via injection. The increase in S only led participants to infer an increase in P and X in the cases in which this increase in S was observed, but *not* for the cases in which the increase in S was due to the effects of an injection.

The results of this experiment provide strong evidence that adult participants rely on both observed evidence and knowledge of the underlying causal structure (represented in a causal graph) when generating predictions about hypothetical scenarios. This supports the proposal that human cognition modifies current representations of causal structure in order to predict the outcomes of hypothetical interventions or reason about counterfactual states. Using this method of causal reasoning, we are therefore able to represent, construct, or modify a causal map of any possible world that we are able to imagine: fictional worlds, pretend scenarios, thought experiments, hypotheticals, future states, etc. Engaging in graphical surgery of our causal maps allows us to consider counterfactual possibilities (the "what ifs" and "if onlys" of our psychological lives), and accounts for the fact that the human concept of causation includes causal relationships that hold regardless of our ability to actually carry out a particular intervention. Assessing whether the relationship between X and Y is causal does not depend upon whether an intervention may actually be performed on X, but instead depends upon what *would* happen to Y *if* that the intervention were to be performed (Woodward, 2007).

We propose, therefore, that through the very same method that is used to represent interventions in the actual world, children and adults are able to represent and manipulate the causal relationships that exist within imagined spaces. To do so, the learner simply takes a causal model of the actual world and changes the value of one or several variables according to the counterfactual assumption. This cuts the arrow that connects the variable to its normal causes. As a result, all other causal influences are rendered inoperative in this imagined space. The graph that exists following the imagined intervention represents the causal structure of the possible world that is being entertained.

For example, you may have forgotten to tie your shoelace when you left the house this morning, which led to your tripping and twisting your ankle on your way to school. In order to engage in a counterfactual analysis, you would begin by modeling the actual event, and make a series of inferences about the consequences within this causal model (e.g., you will not be able to go mountain climbing tomorrow). Next, you can model what would have happened had you stopped to tie your shoe before leaving the house this morning. This is done via an imagined intervention that fixes the value of the variable that represents the state of your shoelace, thus altering the outcome of the causal events in the model.

Modeling of counterfactuals therefore requires updating the current causal model at least two times: once to calculate the probabilities of the events conditional on the observed facts, and a second time to reanalyze the probabilities in the mutilated graph. Using this mutilated graph, we may make inferences about novel effects that would likely occur in the possible world that we have generated, and even make further (potentially dramatic) changes to the causal structure through additional imagined interventions. The *do* operator therefore facilitates reasoning about fictional possibilities without necessarily impacting the beliefs that we hold about the actual

world. However, we may choose to incorporate this counterfactual information to our causal knowledge, use it to inform our future decision-making or planning, or apply it to aid in reasoning about the actual state of the world.

According to this interventionist account of causality, children (and adults) are equipped with all of the tools they need to reason about hypothetical interventions and counterfactuals: generating a counterfactual is causally equivalent to engaging in an imaginary intervention on a causal model. No special cognitive resources are required for imaginative acts. Reasoning counterfactually is simply the process of making an assumption, and then following the various causal implications of that assumption to generate some novel pattern of effects. This may be done to generate a near counterfactual, by making a realistic assumption that could have easily been true (e.g., If you had tied your shoelace then you would not have tripped and twisted your ankle.). In other, more distant counterfactuals, an assumption is made that is generally known to be false and the causal implications are traced through to a set of counterfactual effects. This is a familiar task that is often used in the construction of fictions: an author makes a particular assumption about the world (e.g., that robots become sentient; that time travel exists; that your toys come alive when you leave the room) and then follows the implications of that assumption downstream.

Part of the reason that we find these possible worlds so engaging is probably because they utilize our natural ability to reason counterfactually, to play with our causal models. As mentioned previously, one of the highly adaptive benefits in engaging with these imaginary causal structures is that we can intervene on the fictional world without changing our beliefs about the real world: these two worlds are kept distinct. These imaginative activities provide a wealth of useful information about a variety of possible outcomes without actually requiring

interventions on the real world. The information that is gathered through imaginative intervention may then be applied to producing change in the real world.

Imagining interventions and playing with causal maps: Bayes nets in cognitive development (h1)

A growing body of research suggests that children are making and using causal maps of the world, manipulating them to imagine new possibilities, and applying this information to make new inferences and perform new actions on the environment (e.g., Gopnik et al., 2001; Schulz & Gopnik, 2004; Schulz, Bonowitz, Griffiths, 2007; Schulz, Gopnik, Glymour, 2007). One method that has been used to explore this phenomenon is to introduce 3- and 4-year-old children to novel causal events, and see whether they use that knowledge to make predictions, design new interventions, and consider new possibilities, including counterfactuals.

The Blicket Detector (h2)

Many of the early studies investigating causal inference in young children utilized a novel device developed by Gopnik & Sobel (2000). This device – the blicket detector – lights up and plays music whenever a "blicket" is placed on it. Some objects are blickets, and some are not, but external appearance is not an obvious indicator. Child participants watch a series of trials in which one or more objects are placed on the blicket detector and the effect is observed. Participants are then asked to indicate which objects are blickets, use this information to "make the machine go," or demonstrate their knowledge by generalizing to a novel blicket detector. Using this method, a variety of researchers have demonstrated children's ability to correctly infer the causal relationship that exists between the objects and the blicket detector after very few repeated observations (Gopnik et al., 2001; Sobel, et al., 2004; Gopnik & Schulz, 2007).

In one of the first of these studies (Gopnik, et al., 2001), 3- and 4-year-old children were taught that a particular block was a "blicket." These children were then shown that the blicket combined with a non-blicket also made the machine go (see Figure 3). When asked to make the machine go, children selected only the blickets. More tellingly, in a subsequent experiment, when children were asked to make the machine stop, they suggested removing the blicket, even though they had never observed the machine being stopped this way during training. These children were able to use the new causal information to draw the correct inferences about the underlying causal mechanism. This included the counterfactual conclusion that removing the blicket would make the machine stop. Children combined their prior knowledge about physical causality with their newly learned causal knowledge about blickets and detectors, and were able to imagine what might happen if you removed the blicket from the machine.

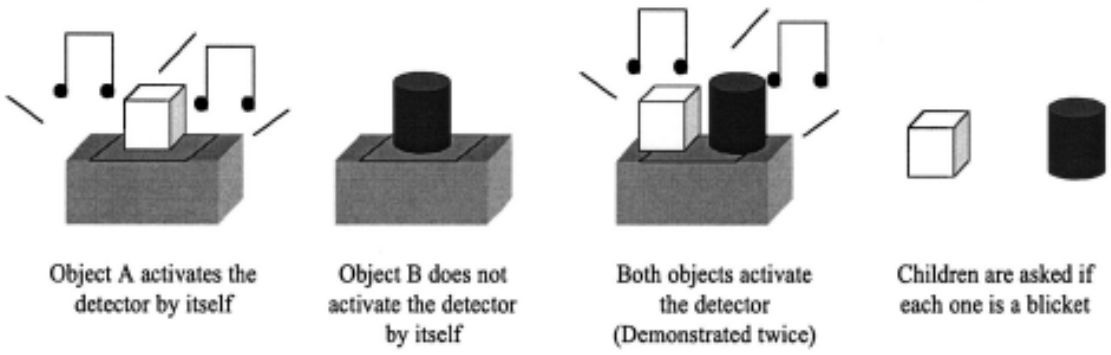

Figure 3. Blicket detector procedure used in Experiment 1 of Gopnik, et al. (2001) (reprinted with permission).

Next, in a series of experiments conducted by Schulz & Gopnik (2004), children were shown a blicket detector with a switch attached. Again, children had no knowledge of how the new machine worked. Children were then asked whether the machine would go when you flipped the switch, or whether it would go if you simply told the machine to go. At pre-test, all children said that the switch could make the machine go, but that speaking to the machine would

not have the same effect. These children had already learned that machines work differently than things with minds. However, after they saw that talking to the machine causes the machine to go, children answered very differently. When asked to make the machine stop, children told the machine to stop, instead of flipping the switch. Further, when these children were asked to predict what would make a new machine go, they were much more likely to suggest talking to the machine as a possible cause than before they had observed this causal relationship. By providing children with new causal knowledge, we can change the possibilities that they will spontaneously entertain, and change the types of actions that they will take**.**

By the time children are 4-years-old, they engage in far more complex experiments with the causal world. In an experiment conducted by Schulz et al (2007), 4-year-olds were introduced to a different novel toy: a box, with two, interconnected gears on top and a switch on the side. When you flip the switch, the gears both turn together. This observation alone is not enough to provide information about how the toy works. By removing one of the gears however, it becomes clear that the underlying causal structure of the gear toy is a causal chain: flipping the switch turns gear B, but not gear A; gear B is responsible for causing the movement of gear A. It is only by exploring the toy that the child would be able to differentiate between a causal chain structure and a common-cause structure.

After observing the initial demonstration, four-year-olds were instructed to determine how the toy works, and were then left alone with the toy. These children played with the gear toy, and broadly explored the box, gears, and switch. While all children engaged in a large number of non-informative actions, many of the children were also able to solve the problem in the context of their free play. Children were equally good at learning all of the causal structures that were presented, and in all cases, when children were shown the appropriate evidence, they

chose the correct structure more often than the other structures. Furthermore, like adults, children were able to predict the outcomes of hypothetical interventions for all the underlying structures (Schulz, Gopnik & Glymour, 2007). They could tell you for example what would happen if the gear A made gear B go, and you removed gear A. These studies provide evidence that as children intervene on the world and observe a range of interventions performed by other people, they are able to infer a variety of different causal structures from the patterns of evidence. They can even make predictions regarding fictitious sets of evidence from their knowledge of causal structure.

Another critical connection between causal knowledge and imagining new possible ways that the world could work is our ability to infer the existence of unobserved or invisible causes that underlie events that we observe. In order to examine this ability through the lens of the causal Bayes net formalism, children were introduced to yet another novel experimental apparatus called a "stickball machine" (Kushnir, Gopnik, Schulz, & Danks, 2003; Schaefer & Gopnik, 2003). This machine operated by moving two stickballs up and down either independently or simultaneously without a visible mechanism (see Figure 4). The experimenter could also visibly intervene on the machine by pulling on the sticks in view of the child participant. This machine was used to test whether children would infer the presence of an unseen cause when interventions on either stickball failed to appropriately correlate with the movement of the other. To do this, children were shown that the movements of stickballs A and B were probabilistically correlated with one another. Participants then saw that pulling up on A did not move B, and that pulling up on B did not move A. According to the Bayes net formalism, if the movements of A and B are dependent upon one another, but intervening on A (*do*[A]) failed to increase the probability of B (and vice versa), then the learner should infer the

existence of an unobserved common cause. This is precisely what children did, suggesting that some invisible mechanism was causing the movement to occur. This type of causal learning process may therefore begin to explain how it is that children and adults are easily able to imagine novel theoretical causes (including magical entities) when available data fails to provide conclusive explanations for phenomena in the world.

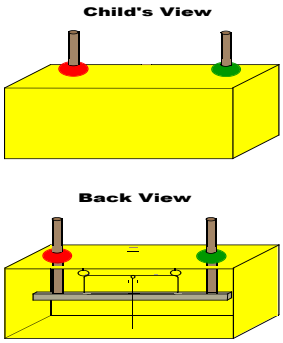

Figure 4. The "stickball machine" used in Kushnir, et al. (2003) (reprinted with permission)

Causation and counterfactuals in childhood (h2)

In the studies presented above, children are using the causal structure of the world to generate one particular kind of counterfactual, namely, a counterfactual that involves considering possible interventions on the world and their consequences. Children develop causal theories of the world from a very early age, and that knowledge allows them to actively intervene on the world. But what about more classic "backwards counterfactuals" – counterfactuals about what might have happened in the past, rather than about what might happen in the future? If causal knowledge and counterfactual thinking go together, then this might explain how young children have the parallel ability to generate backward counterfactuals and to explore possible worlds that do not exist.

Developmental psychologists have begun to examine children's understanding of this link between causal and counterfactual reasoning. In early work, Harris, German, and Mills (1996) conducted a series of studies in which 3- and 4-year-old children were presented with scenarios in which they were asked to reason about a causal sequence (e.g., a character walks across the floor with muddy boots making a mess of the floor). These children were asked counterfactual questions about what would have happened had the events occurred differently. Children were able to answer correctly, demonstrating their early ability to appreciate the link between causal and counterfactual claims, and employ counterfactuals in everyday reasoning. In a subsequent experiment, Harris, et al. (1996) presented children with scenarios in which altering the antecedent would prevent an accident from occurring (e.g., the character's fingers being covered with ink). Again, children were able to offer appropriate responses regarding why the accident occurred, and what could have been done to prevent it (e.g., using a pencil instead of a pen). As these examples show, children are able to evaluate fictional information in light of their background knowledge regarding the causal structure of the world, and rely on inferential processes to update this information accordingly.

In later work, Sobel and Gopnik (2003) demonstrated that children who are able to make correct predictions regarding hypothetical future scenarios were also able to reason correctly about counterfactual claims in the same domain. This provided some evidence that once children have an accurate causal map (an understanding of causal structure in a particular domain), they are able to make predictions and engage in counterfactual reasoning. In a later study, Sobel (2004) compared children's ability to generate explanations of events in a particular domain with their ability to consider counterfactuals that would produce these events. To test this, 3- and 4-year-old children were presented with stories involving possible and impossible

events (events that violated underlying causal principles) across physical, biological, and psychological domains. They were then asked to provide an alternative action that could have been taken to result in the occurrence of each of the possible and impossible events for the three domains.

Results of this study demonstrated that children are able to provide explanations for impossible events, and even generate appropriate counterfactuals for the possible events. While younger children experienced some difficulty stating that there was no counterfactual alternative for impossible events, results did indicate a strong relationship between children's ability to explain impossible events in each domain and their accurate perception that no counterfactual could be generated for these impossible events. In other words, success in producing domain-specific explanations was correlated with successful recognition that counterfactuals cannot be generated to produce impossible events. Sobel (2004) concluded that the maturity of children's causal knowledge in a domain influences their fluency with counterfactual reasoning, and the types of counterfactuals that they will tend to produce. Other research that has been conducted examining children's causal explanations in a variety of domains has provided additional support for this claim, demonstrating that the knowledge structures that children employ in generating causal explanations are coherent and support counterfactuals (Gopnik & Melt off, 1997; Gopnik & Wellman, 1994; Wellman & Liu, 2007).

The implication of these findings is related to the proposal that adults do not tend to propose counterfactual scenarios that violate physical laws (Cheng & Novice, 1992; Harris, 2000; Selah, et al., 1995). For example, adults who are told about a plane crash do not spontaneously consider how this outcome may have been different had gravity suddenly discontinued to have the same effect. These far counterfactuals tend to be automatically rejected

during counterfactual reasoning about reality. Thinking counterfactually may therefore depend on the way that children and adults represent their causal knowledge in a given domain. This contrasts with prevailing accounts of counterfactual reasoning that argue that it is a general ability associated with children's developing understanding of mental representation (Guajardo & Turley-Ames, 2001; Riggs, Peterson, Robinson, & Mitchell, 1998). Instead, children's developing understanding of causal structure in each domain of knowledge influences the counterfactuals that they consider and spontaneously generate (Sobel, 2004). This link between causal knowledge and imagination might also explain the cases where children *do not* think counterfactually. Children might sometimes fail to think counterfactually because they don't have the right kind of causal knowledge, not because they're unable to imagine possibilities, just as it is difficult for most adults to explain what could have been done to prevent the space shuttle crash, or what should be done to prevent it in the future.

Pretense and Causality (h2)

While the role of pretend play in cognitive development is not well understood, the link between the ability to think counterfactually and causal knowledge may underlie young children's engagement in pretense activities as well. The majority of research examining pretend play has focused on the link between pretense and theory of mind (Leslie, 1987; 1994; Lillard, 1993; Nichols & Stich, 2003). Because pretend actions have clear parallels with actions that are based upon false belief, and because pretend play depends on the child's ability to decouple from reality to consider alternative states, it has long been assumed that pretend play is a product of the development of theory of mind (Leslie, 1987).

However, another way to view pretend play is to consider theory of mind as just one facet of a broader category of causal reasoning. From 2- to 6-years of age, children discover facts

about how their own minds and the minds of others work: they formulate a causal map of the mind. They start to understand the causal connections between desires and beliefs, emotions and actions, just as they start to understand the connections between blickets and blicket detectors. According to this more general picture, pretend play reflects possible ways that the world might be. In particular, both learning and reasoning in the probabilistic models framework require the ability to generate patterns of evidence from an initially false premise -- the basic cognitive ability behind pretense. In Bayesian learning, children consider an alternative causal model that initially has a lower probability than the model they currently adopt. Then they must generate a sample of evidence from that alternative model and compare it to the current "true" model and the current evidence. If the new model was more likely to generate the observed evidence than the old model, it will replace that model and become the new "true" model. In reasoning about interventions, children must similarly mentally "set" a variable to a new value and then consider the down-stream consequences of that change.

Pretend play may be therefore understood not as a product of theory of mind, but as a precocious display of children's developing abilities in counterfactual reasoning – setting false premises (assuming a possible world in which there is "tea" inside the empty cup), and following the effects of this counterfactual premise downstream. We are able to modify our existing causal maps without changing our beliefs about the causal structure of the world, and this capacity both supports pretend play and allows children to exercise this essential cognitive tool. Therefore, in much the same way that exploratory play allows children to discover the causal structure of the physical world, pretend play likely facilitates the early development of counterfactual reasoning abilities – which are essential to the probabilistic model of causal learning and to Bayesian learning. From this perspective, not only does causation give fantasy its logic (young children

are quite proficient at tracking the causal rules in fictional worlds) (e.g., Harris, 2000; Onishi, Baillargeon, & Leslie, 2007; Skolnick & Bloom, 2006), but the act of engaging in fantasy helps causal learning to occur. According to this model, pretend play may be interpreted as an engine of learning in development, in which children are able to practice their developing skills in reasoning counterfactually about the causal world in a variety of domains.

This relationship has only recently begun to be empirically explored. Buchsbaum & Gopnik (unpublished data) have been working on testing the hypothesis that pretense may act as a form of counterfactual causal reasoning, allowing children to explore causal "what if" scenarios in imaginary spaces. In one study, preschool children were taught a novel causal relationship, in which one object (a "zando") activates a blicket detector machine, while another object (a "non-zando") does nothing. After learning these relationships, children were asked two explicit counterfactual questions: (1) "If the zando were not a zando, would the machine play music?" and (2) "If the non-zando were a zando, would the machine play music?" The majority of children answered both counterfactual questions correctly. In the second phase of the experiment, children were introduced to a box and two blocks. They were asked to pretend that the box was the machine and that one block was a zando, while the other block was a non-zando. After placing each block on the pretend machine, children were asked whether there was music playing in the pretense activity.

Children's causal inferences about the pretend objects were generally consistent with the objects' real causal roles, demonstrating children's ability to maintain newly learned causal relationships within the context of a pretense scenario. More tellingly, the children who answered the counterfactual questions incorrectly were much more likely to answer the pretense questions incorrectly. These children were attending to the true identity of the blocks; they were

less likely to say that there was pretend music when the pretend zando was placed on the machine. These preliminary results do seem to support the link between counterfactual reasoning abilities and engagement in causal pretense, and this relationship is currently being further explored in our lab.

Along these same lines, children's developing ability to think counterfactually and reason about causality may also underlie why young children create imaginary companions. Imaginary companions may be construed as an elaborate example of psychological counterfactuals; they reflect possible ways that people might be, and possible ways they might act in the world under a variety of constructed circumstances. This relationship explains Taylor's (1999) findings that children with imaginary companions tend to have a more advanced theory of mind than other children of the same age, despite similarities in overall intelligence. Further, the shift from imaginary companions in early childhood to the "paracosms" that are more commonly observed in late childhood may also reflect shifts in children's causal knowledge of other people. Older children, who already understand how individual minds work, become more interested in what happens when minds interact in socially complex ways. These older children are no longer primarily interested in understanding individual people. Instead, they are trying to understand the elaborate social networks that will be crucial for their adult lives. Paracosms are a way of exploring counterfactual societies, just as imaginary companions are a way of exploring counterfactual minds (Gopnik, 2009).

We propose therefore, that children pretend so much because they are learning so much. Our causal maps allow us both to understand the existing physical and psychological world and to invent and realize new physical and psychological worlds. The line between a fiction and a close counterfactual is one of degree rather than kind: fictions are counterfactuals that just

happen to be further away from our real world than other possible worlds. It is possible that by engaging in far counterfactuals, it allows children to fill in parts of their developing causal maps that are not accessible through real interventions and observations. By exploring the far boundaries of the possibility space – causal events with particularly low probabilities – children may be acting to highlight the areas that lie outside the boundaries of their limited experience in the real world.

Conclusion (h1)

In this chapter, we have outlined the Bayesian picture of causality that has had a large impact on theories about human learning about the real world. However, this picture of causality also carries far-reaching implications for the study of the imagination. The Bayesian theory of learning depends upon the idea that children can imagine alternatives to their current causal maps of the world. Children construct alternative hypotheses about what the world is like, and they compare and contrast different possible causal maps. As children's theories change over the course of development, and their causal maps become more accurate, with parameters that more closely approximate the events in the world, the counterfactuals that they are able to produce and the possibilities that they are able to envision become richer. These counterfactuals allow children to create different worlds, and even intervene to make some of those alternatives real. In this way, imagination can be interpreted to depend in large part on causal knowledge; in the same way that developing causal knowledge depends upon the ability to imagine.

On the probabilistic model view when children represent causal structure, they have all the necessary cognitive tools for reasoning about counterfactuals and for representing imaginary worlds. According to these ideas about the process underlying human learning, there are no special cognitive resources that are required to explain imaginative acts. Instead, reasoning

counterfactually is simply the process of making an assumption, and then following the various causal implications of that assumption to generate some novel pattern of effects. Fictional worlds that are less similar to the real world, or further away in the space of possibilities, simply contain fewer causal relations that are also true of the real world. However, by relying upon our inference systems, we are able to draw true conclusions from these false premises, and use this information to inform our reasoning about the real world. When we engage in an imaginative act, we simply produce a counterfactual premise, and proceed from there – generating the causal consequences that would be appropriate if the premise were indeed true. In the space of possible causal graphs, the imaginative world has a low probability, but follows the same causal laws and has the same relationship to the data.

This theory of causal reasoning provides an account of the cognitive mechanism that may underlie this process of inferential elaboration. According to this interpretation of the imagination, reasoning about possible worlds is no different from reasoning about the actual world – and in fact, with the important exception of the false premises upon which the counterfactual is based, the causal structure of the imaginary space *does* tend to conform to the structure in the real world (e.g., Onishi, Baillargeon, & Leslie, 2007; Harris, 2000; Harris & Kavanaugh, 1993; Weisberg & Goodstein, 2009). By incorporating the imaginary realm, we are able to learn about the causal structure of the world without necessarily acting in the physical world. In this way, learning the truth about the world and creating new worlds represent two sides of the same coin. Causal theories in childhood tell us both what is true *and* what is possible.

Future directions

1. Considering the amount of information that children are exposed to in the form of fictional stories and imaginary representations of the world, it will be important to understand the inferences that underlie children's causal learning in this domain. To begin to address this issue, Walker, Ganea, & Gopnik (unpublished data) are currently exploring the following questions: When do children choose to transfer causal information from the fictional space to the real world, and do contextual cues influence generalization to the real world? Does the likelihood that children will generalize causal structure from the fictional to the real world vary based upon the perceived proximity of the causal map supporting the fictional world to the causal map of the real world? Does this sensitivity to the similarity between causal maps change over the course of development? Does the process of learning from fictional information fit into a Bayesian framework?
2. What is precise the nature of the relationship that exists between counterfactual reasoning and pretend play?
3. Are there certain causal structures that are impossible to learn without engaging with far counterfactuals? Do we use far counterfactuals to explore the boundary of the possibility space that would be otherwise inaccessible to us (e.g., the possible worlds that are constructed for the purposes of thought experimentation)?
4. How can the Bayesian account of causality be used to explain the phenomenon of imaginative resistance? Are there certain facts that are so causally central to the representation of the world that they simply cannot be counterfactualized?

5. What is the role of immaturity for imaginative activities during development? Is there a correlation between lack of inhibition in attentional processes and the tendency to more broadly explore the space of possible counterfactuals?

Endnotes

[1] This is not to say that children have conscious awareness of these graphical representations or Bayesian learning procedures. Instead, this proposal attempts to outline causal learning at the computational level (see Marr, 1982 for an explanation of the levels of analysis in scientific explanation).

References

Baker, C., Saxe, R., & Tenenbaum, J.B. (2006). Bayesian models of human action understanding. In Y. Weiss, B. Scholkopf, & J. Platt (Eds.), *Advances in Neural Information Processing Systems,* 18: 99-106.

Bullock, M., Gelman, R., & Baillargeon, R. (1982). The development of causal reasoning. In W.J. Freidman (Ed.), *The developmental psychology of time* (pp. 209-254). New York: Academic Press.

Campbell, J. (1994). *Past, space, and self*. Cambridge, Mass.: MIT Press.

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

Chater, N. & Manning, C.D. (2006). Probabilistic models of cognition: conceptual foundations. *Trends in Cognitive Sciences,* 10: 287-291.

Chater, N., Tenenbaum, J., Yuille, A.L. (2006). Probabilistic models of cognition: Where next? In *Trends in Cognitive Neuroscience*, 10(7): 292-293.

Cheng, P. W. & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review, 99,* 365-382.

Downman, M. (2002). Modeling the acquisition of colour words. In *Proceedings of the 15th Australian Joint Conference on Artificial Intelligence: Advances in Artificial Intelligence,* 259-271. Springer-Verlag.

Gelman, S. A., & Wellman, H. M. (1991). Insides and essence: Early understandings of the non-obvious. *Cognition, 38*(3), 213-244.

Glymour, C. (2001). *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*. MIT Press: Cambridge, MA.

Glymour, C. (2003). Learning, prediction and causal Bayes nets. *Trends in Cognitive Science, 7*, 43-48.

Goodman, N.D., Baker, C.L., Bonawitz, E.B., Mansinghka, V.K., Gopnik, A., Wellman, H., Schulz, L., & Tenenbaum, J.B. (2006). Intuitive theories of mind: a rational approach to false belief. *Proceedings of the 28th Annual Conference of the Cognitive Science Society.* Mahway, NJ: Erlbaum.

Gopnik, A. (1988). Conceptual and semantic development as theory change. *Mind and Language* 3, (Autumn): 197-216.

Gopnik, A. (2000). Explanation as orgasm and the drive for causal understanding: The evolution, function, and phenomenology of the theory-formation system. In F. Keil & R. Wilson (Eds.) *Cognition and explanation.* Cambridge, MA: MIT Press (299-323).

Gopnik, A. (2009). *The Philosophical Baby*. Farrar, Straus, & Giroux: New York.

Gopnik, A. Glymour, C., Sobel, D.M., Schulz, L.E., Kushnir, T., and Danks, D. 2004. A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review* 111, (1) (Jan.): 3-32.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts and theories*. Cambridge, MA: MIT Press.

Gopnik, A. & Sobel, D.M. (2000). Detecting blickets: How young children use information about causal properties in categorization and induction. *Child Development*, 71: 1205-1222.

Gopnik, A., Sobel, D.M., Schulz, L.E., and 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, (5) (Sept.): 620-9.

Gopnik, A. & Tenenbaum, J.B. (2007). Bayesian networks, Bayesian learning, and cognitive development*, Developmental Science*, 10(3): 281-287.

Gopnik, A., & Wellman, H. M. (1994). The theory theory. In S. A. Gelman & L. A. Hirschfeld (Eds.), *Mapping the mind: Domain specificity in cognition and culture; Based on a conference entitled "Cultural Knowledge and Domain Specificity," held in Ann Arbor, MI, Oct 13-16, 1990* (pp. 257-293). New York, NY, US: Cambridge University Press.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, 51: 354-384.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14 (8): 357-364.

Guajardo, N.R. & Turley-Ames, K.J. (2001). *Theory of mind and counterfactual thinking: Mutating the antecedent versus the consequent.* Poster presented at the biennial meeting of the Society for Research in Child Development, Minneapolis.

Hagmayer, Y., Sloman, S., Lagnado, D., & Waldmann, M.R. (2007). Causal reasoning though intervention, in *Causal Learning: Psychology, Philosophy, & Computation*, A. Gopnik & L. Schulz (Eds.). Oxford University Press: Oxford.

Harris, P.L. (2000). *The work of the imagination*. Malden, Mass.: Blackwell Publishing.

Harris, P.L., German, T. and Mills, P. (1996). Children's use of counterfactual thinking in causal reasoning. *Cognition* 61, (3) (Dec.): 233-59.

Harris, P.L. & Kavanaugh, R.D. (1993). Young children's understanding of pretense. *Society for Research in Cognitive Development Monographs* (231).

Hickling, A.K., and Wellman, H.M. (2001). The emergence of children's causal explanations and theories: Evidence from everyday conversation. *Developmental Psychology* 37, (5) (Sept.): 668-83.

Hume, D. (2007). *An enquiry concerning human understanding*. Oxford world's classics. Oxford, England; New York: Oxford University Press.

Inagaki, K., and Hatano, G. (2006). Young children's conception of the biological world. *Current Directions in Psychological Science* 15, (4) (Aug.): 177-81.

Kushnir, T., Gopnik, A., Schulz, L., & Danks, D. (2003). *Inferring hidden causes.* Paper presented at the 25th Conference of the Cognitive Science Society.

Lagnado, D. A, & Sloman, S. A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 856-876.

Legare, C.H., Gelman, S.A., & Wellman, H.M. (2010). Inconsistency with prior knowledge triggers children's causal explanatory reasoning. *Child Development,* 81: 929-944.

Leslie, A.M. (1987). Pretense and representation: The origins of "theory of mind." *Psychological Review* 94, (4) (Oct.): 412-26.

Lewis, D. (1986). *Counterfactuals*. Cambridge, Mass.: Harvard University Press.

Lillard, A.S. (1993). Young children's conceptualization of pretense: Action or mental representational state? *Child Development,* 64: 372-386.

Lu, H., Yuille, A., Liljeholm, M., Cheng, P.W., & Holyoak, K.J. (2006). Modeling causal learning using Bayesian generic priors on generative and preventive powers. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society*: 519-524.

Lucas, C., Gopnik, A., & Griffiths, T.L. (2010). Developmental differences in learning the forms of causal relationships, *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.

Lucas, C. & Griffiths, T.L. (2010). Learning the form of causal relationships using Hierarchical Bayesian Models, *Cognitive Science*, 34: 113-147.

Mackie, J.L. (1974). Truth, Probability, and Paradox a Reply to James E. Tomberlin's Review. *Philosophy and Phenomenological Research* 34 (4): 593-594.

Marr, D. (1982). *Vision: A computational investigation in the human representation and processing of visual information.* San Francisco: Freeman.

Nichols, S. & Stich, S. (2000). A cognitive theory of pretense. *Cognition,* 74: 115-147.

Niyogi, S. (2002). Bayesian learning at the syntax-semantics interface. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (W. Gray & C. Schunn, Eds., 697-702). Mahwah, NJ: Erlbaum.

O'Keefe, J. and Nadel, L. 1979. Précis of O'Keefe and Nadel's The hippocampus as a cognitive map. *Behavioral and Brain Sciences* 2, (4) (Dec.): 487-533.

Onishi, K.H., Baillargeon, R., & Leslie, A.M. (2007). 15-month-old infants detect violations in pretend scenarios. *Acta Psychologica*, 124(1): 106-128.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. New York: Cambridge University Press.

Perner, J. (1991). *Understanding the representational mind.* Cambridge, MA: MIT Press.

Piaget, J. (1929). *The child's conception of the world*. New York: Harcourt, Brace.

Regier, T. & Gahl, S. (2004). Learning the unlearnable: the role of missing evidence. *Cognition*, 93: 147-155.

Riggs, K.J., Peterson, D.M., Robinson, E.J., & Mitchell, P. (1998). Are errors in false belief tasks symptomatic of a broader difficulty with counterfactuality? *Cognitive Development,* 13: 73-91.

Schulz, L. (2003). *The play's the thing: Interventions and causal inference.* Paper presented at the biennial meeting of the Society for Research in Child Development, Tampa, FL.

Schulz, L.E., Gopnik, A., and Glymour, C. (2007). Preschool children learn about causal structure from conditional interventions. *Developmental Science* 10, (3) (May): 322-32.

Schulz, L., Bonawitz, E.B., & Griffiths, T.L. (2007). Can being scared give you a tummy ache? Naïve theories, ambiguous evidence, and preschoolers' causal inference, *Developmental Psychology,* 43(5): 1124-1139.

Schulz, L.E., Kushnir, T., & Gopnik, A. (2007). Learning from doing: Interventions and Causal Inference, in *Causal Learning: Psychology, Philosophy, & Computation*, A. Gopnik & L. Schulz (Eds.). Oxford University Press: Oxford.

Seelau, E.P., Seelau, S.M., Wells, G.L., and Windschitl, P.D. (1995). Counterfactual constraints. In *What might have been: The social psychology of counterfactual thinking,* N.J. Roese & J.M. Olson (Eds.). Mahwah, NJ: Lawrence Erlbaum Associates.

Schaefer, C. & Gopnik, A. (2003). Causal reasoning in young children: The role of unobserved variables. Poster presented at the biennial meeting of the Society for Research in Child Development.

Shultz, T.R. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development*, 47 (194).

Skolnick, D. & Bloom, P. (2006). What does Batman think about SpongeBob? Children's understanding of the fantasy/fantasy distinction. *Cognition,* 101: B9-B18.

Slaughter, V., Jaakola, R., & Carey, S. (1999). Constructing a coherent theory: Children's biological understanding of life and death. In M. Siegel & C. Peterson (Eds.), *Children's understanding of biology and health* (pg. 71-96). Cambridge, MA: Cambridge University Press.

Sloman, S. (2005). *Causal Models.* Oxford University Press: Oxford.

Sloman, S. & Lagnado, D. (2005). Do we "do"? *Cognitive Science*, 29: 5-39.

Sobel, D.M. (2004). Exploring the coherence of young children's explanatory abilities: Evidence from generating counterfactuals. *British Journal of Developmental Psychology* 22, (1) (Mar.): 37-58.

Sobel, D.M. & Gopnik, A. (2003). *Causal prediction and counterfactual reasoning in young children: Separate or similar processes?* Unpublished Manuscript.

Sobel, D.M. & Kushnir, T. (2006). The importance of decision demands in causal learning from interventions, *Memory and Cognition*, 34: 411-419.

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review,* 99(4): 605-632.

Spirtes, P., Glymour, C.N., and Scheines, R. 1993. *Causation, prediction, and search.* New York: Springer-Verlag.

Taylor, M. (1999). *Imaginary companions and the children who create them*. New York: Oxford University Press.

Taylor, M., Hodges, S. D., & Kohanyi, A. (2003). The illusion of independent agency: Do adult fiction writers experience their characters as having minds of their own? *Imagination, Cognition, and Personality,* 22: 361-380.

Tenenbaum, J. B., & Griffiths, T. L. (2003). Theory-based causal induction. In *Advances in neural information processing systems, 15*: 35-42. Cambridge: MIT Press.

Tenenbaum, J.B., Griffiths, T.L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences,* 10: 309-318.

Tenenbaum, J.B. & Xu, F. (2000). Word learning as Bayesian inference. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*: 517-522.

Tolman, E.C. (1948). Cognitive maps in rats and men. *Psychological Review* 55, (4) (July): 189-208.

Waldmann, M. R., & Hagmayer, Y. (2005). Seeing versus doing: Two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Motivation, and Cognition, 31*, 216-227.

Weisberg, D. S., & Goodstein, J. (2009). What Belongs in a Fictional World? *Journal of Cognition and Culture*, *9*(1), 69-78.

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

Wellman, H. M., & Gelman, S. A. (1998). Knowledge acquisition in foundational domains. In W. Damon (Series Ed.) & D. Kuhn & R. Siegler (Vol. Eds.), *Handbook of child psychology: Vol. 2. Cognition, perception, and language* (5th ed., pp. 523-573). New York: Wiley.

Wellman, H.M. & Liu, D. (2007). Causal reasoning as informed by the early development of explanations, in *Causal Learning: Psychology, Philosophy, & Computation*, A. Gopnik & L. Schulz (Eds.). Oxford University Press: Oxford.

Woodward, J. 2003. *Making things happen: A theory of causal explanation.* Oxford: Oxford University Press.

Woodward, J. 2007. Interventionist theories of causation in psychological perspective, in *Causal Learning: Psychology, Philosophy, & Computation*, A. Gopnik & L. Schulz (Eds.). Oxford University Press: Oxford.

Xu, F. & Tenenbaum, J.B. (2007). Word learning as Bayesian inference. *Psychological Review*, 114: 245-292.