Using Physical Theories to Infer Hidden Causal Structure

Thomas L. Griffiths
gruffydd@psych.stanford.edu
Department of Psychology
Stanford University

Elizabeth R. Baraff & Joshua B. Tenenbaum
{liz_b,jbt}@mit.edu
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

Abstract

We argue that human judgments about hidden causal structure can be explained as the operation of domain-general statistical inference over causal models constructed using domain knowledge. We present Bayesian models of causal induction in two previous experiments and a new study. Hypothetical causal models are generated by theories expressing two essential aspects of abstract knowledge about causal mechanisms: which causal relations are plausible, and what functional form they take.

Everyday reasoning draws on notions that go far beyond the observable world, just as modern science draws upon theoretical constructs beyond the limits of measurement. The richness of our naive theories is a direct result of our ability to postulate hidden causal structure. This capacity to reason about unobserved causes forms an essential part of cognition from early in life, whether we are reasoning about the forces involved in physical systems (e.g., Shultz, 1982), the mental states of others (e.g., Perner, 1991), or the essential properties of natural kinds (e.g., Gelman & Wellman, 1991).

The central role of hidden causes in naive theories makes the question of how people infer hidden causal structure fundamental to understanding human reasoning. Psychological research has shown that people can infer the existence of hidden causes from otherwise unexplained events (Ahn & Luhmann, 2003), and determine hidden causal structure from very little data (Kushnir, Gopnik, Schulz, & Danks, 2003). This work has parallels in computer science, where the development of a formalism for reasoning about causality – causal graphical models – has led to algorithms that use patterns of dependency to identify causal relationships (Pearl, 2000; Spirtes, Glymour, & Scheines, 1993). It has recently been proposed, chiefly by Gopnik, Glymour, and their colleagues (Glymour, 2001; Gopnik, Glymour, Sobel, Schulz, Kushnir, & Danks, 2004), that these algorithms may also explain how people infer causal structure.

A fundamental issue in explaining how people infer causal relationships is accounting for the interaction between abstract causal knowledge and statistical inference. The classic debate between approaches that emphasize cause-effect covariation and those that emphasize mechanism knowledge (e.g., Newsome, 2003) turns on this issue. Causal graphical models provide a language in which the problem of causal induction can be formally expressed. However, conventional algorithms for inducing causal structure (e.g., Pearl, 2000; Spirtes et al., 1993) do not provide a satisfying account of either the roles of causal knowledge or statistical inference, or their interaction. These algorithms use tests of statistical independence to establish constraints that must be satisfied by causal structures consistent with the observed data. No knowledge of how causal mechanisms operate, or the functional form of relationships between cause and effect, enters into the inference process. As we argue below, such knowledge is necessary to explain how people are able to infer causal structure from very small samples, and to infer hidden causes from purely observational data. Constraint-based methods are also unable to explain people's graded sensitivity to the strength of evidence for a causal structure, because they reason deductively from constraints to consistent structures.

We will present a rational account of human inference, Theory-Based Causal Induction, which emphasizes the interaction between causal knowledge and statistical learning. Causal knowledge appears in the form of causal theories, specifying the principles by which causal relationships operate in a given domain. These theories are used to generate hypothesis spaces of causal models – some with hidden causes, some without – that can be evaluated by domain-general statistical inference. We will use this framework to develop models of people's inferences about hidden causes in two physical systems: a mechanical system called the stick-ball machine (Kushnir et al., 2003), and a dynamical system involving an explosive compound called Nitro X.

Theory-based causal induction

Our account of causal induction builds on causal graphical models, extending the formalism to incorporate the abstract knowledge about causal mechanism that plays an essential role in human inferences. We will briefly introduce causal graphical models, consider how prior knowledge influences causal induction, and describe how we formalize the contribution of causal theories.

Causal graphical models

Graphical models represent the dependency structure of a joint probability distribution using a graph in which nodes are variables and edges indicate dependence. The graphical structure supports efficient computation of the probabilities of events involving these variables. In a *causal* graphical model the edges indicate causal dependencies, with the direction of the arrow indicating the direction of causation, and they support inferences about the effects of interventions (Pearl, 2000). An intervention is an event in which a variable is forced to hold a value, independent of any other variables on which it might depend. Intervention on a variable A is denoted do(A). Probabilistic inference on a modifified graph, in which incoming edges to A are removed, can be used to assess the consequences of intervening on A.

The structure of a causal graphical model implies a pattern of dependency among variables under observation and intervention. Conventional algorithms for inferring causal structure use standard statistical tests, such as Pearson's χ^2 test, to find the pattern of dependencies among variables, and then deductively identify the structure(s) consistent with that pattern (e.g., Spirtes et al., 1993). These "constraint-based" algorithms can also exploit the results of interventions, and often require both observations and interventions in order to identify the hidden causal structure. Gopnik, Glymour, and colleagues have suggested that this kind of constraint-based reasoning may underlie human causal induction (Glymour, 2001; Gopnik et al., 2004; Kushnir et al., 2003).

The role of causal theories

Constraint-based algorithms for causal induction make relatively little use of prior knowledge. While particular causal relationships can be ruled out a priori, there is no way to represent the belief that one structure may be more likely than another. Furthermore, the use of statistical tests like χ^2 makes only weak assumptions about the form of causal relationships: these tests simply assess dependency, regardless of whether a relationship is positive or negative, deterministic or probabilistic, strong or weak.

Several researchers (e.g., Shultz, 1982) have argued that knowledge of causal mechanism plays a central role in human causal induction. Mechanism knowledge is usually cited in arguments against statistical causal induction, but we view it as critical to explaining how statistical inferences about causal structure are possible from sparse data. Knowledge about causal mechanisms provides two kinds of restrictions on possible causal models: restrictions on which relationships are plausible, and restrictions on the functional form of those relationships. Restrictions on plausibility might indicate that one causal structure is more likely than another, while restrictions on functional form might indicate that a particular relationship should be positive and strong.

These restrictions have important implications for causal induction algorithms. If all structures are possible, both observations and interventions are typically required to identify hidden causes, and without strong assumptions about the functional form of causal relationships, samples must be relatively large. With limitations on the set of possible causal structures and expectations about functional form, however, it is possible to make causal inferences from just observations and from small samples – important properties of human causal induction.

Using causal theories in causal induction

The causal mechanism knowledge that is relevant for statistical causal inference may be quite abstract, and may also vary across domains. Much of this knowledge may be represented in intuitive domain theories. In contrast to Gopnik et al. (2004), who suggest that causal graphical models are the primary substrate for intuitive theories, we emphasize the role of intuitive theories at a more abstract level, providing restrictions on the set of causal models under consideration. Such restrictions cannot be represented as part of a causal graphical model: causal graphical models express the relations that hold among a finite set of propositions, while causal theories involve statements about all relations that could hold among entities in a given domain.

Formally, we view causal theories as *hypothesis space generators*: a theory is a set of principles that can be used to generate a hypothesis space of causal models, which are compared via Bayesian inference. The principles that comprise a theory specify which relations are plausible and the functional form of those relations. These principles articulate how causal relationships operate in a given domain, but need not identify the mechanisms underlying such relationships: all that is necessary for causal induction is the possibility that some mechanism exists, and expectations about the functional form associated with that mechanism. This vague and abstract mechanism knowledge is consistent with the finding that people's understanding of causal mechanism is surprisingly shallow (Rozenblit & Keil, 2002).

In the remainder of the paper, we will demonstrate how Theory-Based Causal Induction can be used to explain human inferences about hidden causes in physical systems. Different systems require different causal theories. We will examine inferences in a mechanical system, the stick-ball machine (Kushnir et al., 2003), and in a dynamical system, Nitro X, which we explore in a new experiment. When reasoning about these systems, people infer hidden causal structure from very few observations, and are sensitive to graded degrees of evidence.

The stick-ball machine

Kushnir et al. (2003) conducted two experiments in which participants had to infer the causal structure

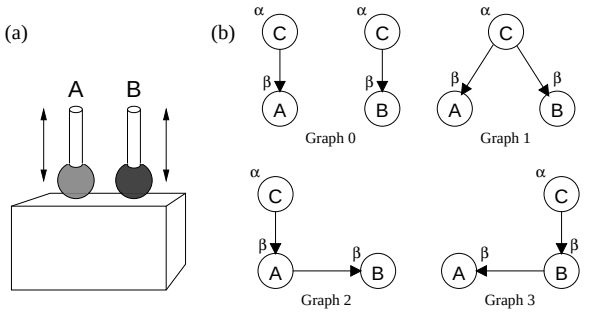

Figure 1: (a) A stick-ball machine. (b) Graphs indicating potential causal structures for the stick-ball machine. Nodes A and B correspond to the two balls, nodes marked C are hidden causes.

of a physical system, the "stick-ball machine", consisting of two colored balls (A and B) mounted on sticks which could move up and down on a box (see Figure 1(a)). The mechanical apparatus moving the balls was concealed, keeping the actual causal relationship unknown. In both experiments, all participants were familiarized with the machine, and told that if one ball caused the other to move it did so "almost always". This probabilistic causal relation was demonstrated by showing the two balls move together four times, an event we denote $4AB$, and A moving alone twice, $2A\bar{B}$. There were three test conditions in Experiment 1, seen by all participants. In the *common unobserved cause* condition, participants saw $4AB$, and four trials in which the experimenter intervened, twice moving A with no effect on B, $2\bar{B}|\text{do}(A)$, and twice moving B with no effect on A, $2\bar{A}|\text{do}(B)$. In the *independent unobserved cause* condition, participants saw $2A\bar{B}$, $2\bar{A}B$, $1AB$, $2\bar{A}|\text{do}(B)$, and $2\bar{B}|\text{do}(A)$. In the *one observed cause* condition, participants saw $4B|\text{do}(A)$ and $2\bar{B}|\text{do}(A)$. Experiment 2 replicated the *common unobserved cause* condition, and compared this with a *pointing control* condition in which interventions were replaced with observations ($4AB$, $2\bar{A}B$, $2A\bar{B}$). The order of conditions and trials within conditions was randomized across participants. In each condition, participants identified the underlying causal structure by indicating graphs similar to those shown in Figure 1(b). The results of both experiments are combined in Figure 2. One causal structure was chosen by the majority of people in each condition – Graph 1 in the *common unobserved cause* condition, Graph 0 in the *independent unobserved causes* condition, Graph 2 in the *one observed cause* condition, and Graph 0 in the *pointing control*.

The results of these experiments provide two challenges to constraint-based accounts. First, people are able to make inferences from small samples – in many cases, far less data than might be required for all relevant χ^2 tests to yield results consistent with the appropriate causal structure. Second, people's

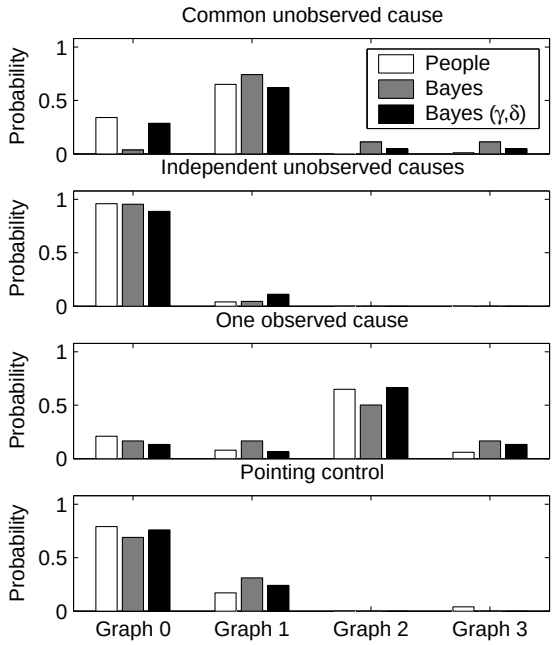

Figure 2: Results of Kushnir et al. (2003), shown with predictions of Bayesian models.

judgments reflect a sensitivity to graded degrees of evidence: in the *independent unobserved causes* condition, over 95% of participants chose Graph 1, while only 60-80% of people chose the most popular structure in the other conditions. This was not simply a consequence of a preference for Graph 0 – the same structure was less popular in the *pointing control* condition, suggesting that there is a difference in the evidence that the data provide for Graph 0 in these two conditions. Constraint-based algorithms are not sensitive to graded degrees of evidence: a causal structure is either consistent or inconsistent with the pattern of dependencies in a dataset.

A theory-based account

Our model of the stick ball machine uses a physical theory that contains three principles:

1. Balls never move without a cause.
2. A hidden cause moves with probability α.
3. A moving cause moves its effect with probability β.

If we add the restrictions that every ball has a single cause and hidden causes never have causes (but can move themselves, per Principle 2), we obtain the four structures shown in Figure 1(b). The principles of the physical theory place strong constraints on the functional form of the causal relationships identified in this structure, allowing us to compute the probability of events involving A and B for each graphical structure, as shown in Table 1.

Given a dataset D, we compute a posterior probability distribution over these structures, $P(\text{Graph } i|D)$, combining prior probabilities,

Table 1: Event probabilities for causal structures

Event	Graph 0	Graph 1	Graph 2
AB	$(\alpha\beta)^2$	$\alpha\beta^2$	$\alpha\beta^2$
$\bar{A}B$	$\alpha\beta(1-\alpha\beta)$	$\alpha\beta(1-\beta)$	0
$A\bar{B}$	$\alpha\beta(1-\alpha\beta)$	$\alpha\beta(1-\beta)$	$\alpha\beta(1-\beta)$
$\bar{A}\bar{B}$	$(1-\alpha\beta)^2$	$1-2\alpha\beta+\alpha\beta^2$	$1-\alpha\beta$
$A\|\text{do}(B)$	$\alpha\beta$	$\alpha\beta$	$\alpha\beta$
$B\|\text{do}(A)$	$\alpha\beta$	$\alpha\beta$	β

Note: Probabilities for Graph 3 are the same as those for Graph 2, exchanging the roles of A and B.

$P(\text{Graph } i)$, with the probability of the observed data under each structure, $P(D|\text{Graph } i)$, using Bayes' rule:

$$P(\text{Graph } i|D) \propto P(D|\text{Graph } i)P(\text{Graph } i)$$

$P(D|\text{Graph } i)$ is the product of the probabilities of the individual events making up D, which can be obtained from Table 1.

If we assume a uniform prior for $P(\text{Graph } i)$, the causal theory leaves two parameters unspecified: α, the probability of a hidden cause moving on a given trial, and β, the probability that a moving cause moves its effect. We set β empirically, via a small experiment. We showed 10 participants a computer simulation of the stick-ball machine, and reproduced the familiarization trials used by Kushnir et al. (2003): participants were told that when A causes B, it makes it move "almost always", and were shown that A moved B on four of six trials. We then asked them how often they expected A would move B. The mean and median response was that A would move B on 75% of trials, so we used $\beta = 0.75$.

Figure 2 shows the predictions of the Bayesian model with $\alpha = 0.47$. The model gave a correlation of $r = 0.94$ with the data, and correctly predicted the most common response in each condition. The model also admits graded degrees of evidence, with the observations and interventions in the *independent unobserved causes* condition providing stronger evidence for Graph 0 than the observations in the *pointing control*. The model departs from people's judgments in one case, failing to predict the minority preference for Graph 0 in the *common unobserved cause* condition. This disparity could have many explanations, such as a default preference for independence between objects, or differences in the salience of different data types and causal structure. For instance, interventions may be weighted higher than observations by a factor of γ, and hidden common causes may receive only a fraction $1/\delta$ of the prior probability accorded to other structures. Figure 2 shows an almost-perfect fit ($r = 0.99$) for such a model, Bayes (γ,δ), with $\gamma = 4, \delta = 2, \alpha = 0.4$. Further experiments will be necessary to determine whether these sorts of psychological variables play a role in the process of causal induction.

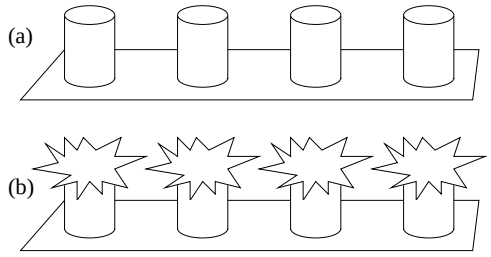

Figure 3: (a) Four cans of the extremely unstable compound Nitro X. (b) A simultaneous explosion.

Nitro X

To provide a further demonstration of the importance of graded degrees of evidence and the ability to infer hidden causes from very little data, we conducted an experiment that tested people's ability to infer the causal structure of a dynamical physical system. Our experiment presents a more severe inductive challenge than the tasks considered by Kushnir et al. (2003), as it requires inferring a hidden common cause from just a single observation, with no verbal cues that such a structure might exist. In the experiment, we introduced people to a novel substance, Nitro X, and illustrated its dynamics: cans of Nitro X could spontaneously explode, and could detonate one another after a time delay that was a linear function of spatial separation, as would be expected from the slow propagation of pressure waves. We then presented them with the *simultaneous* explosion of several cans, without the delays characteristic of pressure waves propagating from one can to the next. We expected that people would see this suspicious coincidence as evidence for some kind of hidden common cause, such as an external force shaking the table. We varied the number of cans, m, to see whether the magnitude of the coincidence had an effect on people's inference to a hidden cause.

Method

Participants Participants were 64 members of the MIT Brain and Cognitive Sciences subject pool, split evenly over four conditions ($m = 2, 3, 4, 6$).

Stimuli The stimuli were pictures of cans sitting on a table, presented on a computer screen. A new set of cans was shown on each trial, and by the end of the trial all cans on the screen had exploded, demonstrated by cartoon explosion graphics like those shown in Figure 3.

Procedure The experiment consisted of three familiarization trials and five test trials. The familiarization trials introduced the participants to Nitro X. In the first trial, participants were told that Nitro X is very unstable, and this was demonstrated by the experimenter tapping a can and the can exploding. In the second trial, participants saw two cans of Nitro X, the experimenter tapped one can,

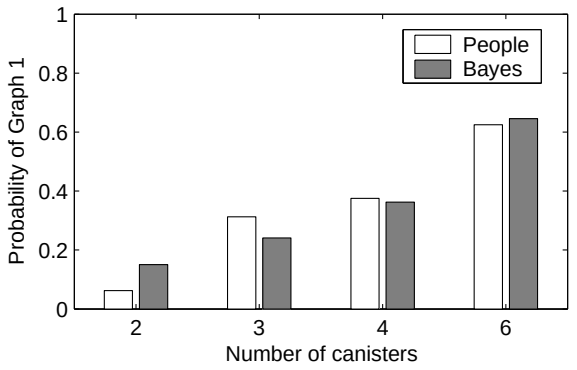

Figure 4: Results of the Nitro X experiment.

which exploded, and the can next to it exploded shortly afterwards. On the third trial, participants were again reminded about the instability of Nitro X, and saw a single can explode without any action by the experimenter, after waiting for a few seconds.

The first two test trials were identical for all four conditions, and both involved four cans exploding in a causal chain, with a delay between successive explosions. In the third test trial, the number of cans in the display was varied, $m = 2, 3, 4$ or 6, depending on condition. After a brief delay, all of the cans exploded simultaneously. The last two test trials allowed the participants to interact with Nitro X by tapping, and will not be discussed further here.

After each test trial, participants were given a sheet of questions for each test trial. These sheets gave three options:

1. The first can exploded spontaneously. That explosion caused the other cans to explode, in a chain reaction.
2. Each can exploded spontaneously, all on its own. There was no causal connection between them.
3. Neither of the above is a likely explanation. Please write a plausible alternative here.

The order of the first two options was counterbalanced, but the third option was always last.

Results and Discussion

For all trials, two rates examined the written responses of participants choosing the third option above, and were in 100% agreement in classifying all such responses as indicating a hidden cause. Over 95% of participants correctly identified the causal chain in the first two trials. The proportion of participants identifying a hidden cause on the third trial, with the simultaneous explosion, is shown in Figure 4. There was a statistically significant effect of m, $\chi^2(3) = 11.36$, $p < 0.01$. The number of cans influenced whether people inferred hidden causal stucture, with most people seeing two cans as independent but six as causally related.

Constraint-based algorithms cannot explain our results. If we imagine that time is broken into discrete intervals, and a can either explodes or does not explode in each interval, then we can construct a contingency table for each pair of cans. Statistical significance tests will identify pairwise dependencies among all cans that explode simultaneously, provided appropriate numbers of non-explosion trials are included. The existence of a hidden common cause is consistent with such a pattern of dependency. However, as a result of reasoning deductively from this pattern, the evidence for such a structure does not increase with m: a hidden common cause is merely consistent with the pattern for all $m > 2$.

This experiment also illustrates that people are willing to infer hidden causal structure from very small samples – just one datapoint – and from observations alone. Constraint-based algorithms cannot solve this problem: while a hidden common cause is consistent with the observed pattern of dependency, causal structures in which the cans influence one another cannot be ruled out without intervention information. People do not consider this possibility because they have learned that the mechanism by which cans influence one another has a time delay. Further situations in which the temporal properties of causal relationships influence causal induction are described by Hagmayer and Waldmann (2002).

A theory-based account

The results of the Nitro X experiment are easy to model: any increasing function of the number of cans would be sufficient. Our goal in modeling these data is to illustrate how Theory-Based Causal Induction extends to a system with non-trivial dynamics and different causal mechanisms, and to show that inferences to hidden causes from the smallest possible sample – a single observation – can have a physically plausible and statistically rational explanation.

We model the explosion times of cans by assuming that at each infinitesimal moment, there is a certain probability that the can will explode. This assumption means that the explosion time of each can follows a Poisson process, with a "rate parameter" determining the probability of explosion at each moment. We set the rates using the following principles:

1. A can explodes spontaneously at rate α.
2. A hidden cause becomes active at rate γ.
3. At the moment a hidden cause is active, a can influenced by that cause explodes at rate $\alpha + \beta$.

A complete theory of Nitro X would need to include further principles stating the functional form of the causal relationship between cans, encoding the fact that this relationship involves a time delay. We have omitted these principles because they do not directly affect the inference to a hidden cause when all explosions are simultaneous.

This theory generates a large number of possible causal structures, with hidden causes influencing various subsets of the cans. We will focus on the two structures shown in Figure 5: Graph 0, in which all

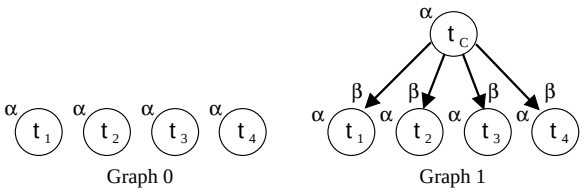

Figure 5: Graphs indicating potential causal structures for the Nitro X experiment.

cans explode spontaneously, is the "null hypothesis" for any inference concerning hidden causes, while Graph 1, in which all cans are also influenced by a hidden cause, gives the highest probability to a simultaneous explosion. These structures are defined on variables representing the time at which cans explode, $t_1, \ldots, t_m$, and the time the hidden cause becomes active, t_C. The inference to a hidden common cause is modeled by computing the posterior probability $P(\text{Graph } 1|T)$, where $T = \{t_1, \ldots, t_m\}$. In a simultaneous explosion, all t_i take the same value, t.

It follows from the theory outlined above that for Graph 0, each t_i is an independent Poisson process with rate α, which gives $P(T|\text{Graph } 0) = \alpha^m \exp\{-m\alpha t\}$. For Graph 1, t_C follows a Poisson process with rate γ. Conditioned on t_C, each t_i is a Poisson process with rate α, except at the moment when the hidden cause becomes active, at which point the rate is $\alpha + \beta$. Computing $P(T|\text{Graph } 1)$ requires integrating over all values of t_C, which we approximate by choosing t_C to maximize $P(T|t_C)$:

$$\begin{aligned} P(T|\text{Graph } 1) &= \int_0^\infty P(T|t_C)P(t_C)\, dt_C \\ &\approx \gamma(\alpha+\beta)^m \exp\{-m\alpha t - \gamma t\} \end{aligned}$$

Applying Bayes' rule, it follows[1] that $P(\text{Graph } 1|T)$ is a sigmoid function of m,

$$P(\text{Graph } 1|T) = \frac{1}{1+\exp\{-gm-b\}}$$

for $g = \log\frac{\alpha+\beta}{\alpha}$ and $b = \log\frac{P(\text{Graph } 1)}{P(\text{Graph } 0)} + \log\gamma - \gamma t$.

The model predicts that increasing m should increase $P(\text{Graph } 1|T)$ for any positive values of α and β, as this results in a positive gain, g. The theory involves four parameters: α, β, γ, and $P(\text{Graph } 0)$. Since these four parameters are not identifiable – multiple sets of parameter values are consistent with the same sigmoid function – we set the parameters of the sigmoid g and b. Using $g = 0.58$ and $b = -2.90$ gives $r = 0.958$, and the predictions shown in Figure 4. These parameters indicate $\beta = 0.79\alpha$ and an initial preference for Graph 0.

Our theory-based approach explains why the number of cans involved in a simultaneous explosion should influence the evidence for a hidden cause, but is clearly not the only model compatible with these data. However, our analysis exposes the rational basis for human judgments, and makes further intuitive predictions that we are in the process of testing. For example, the $-\gamma t$ term in the expression for b indicates that, all other things being equal, decreasing the time before a simultaneous explosion increases the evidence for a hidden cause.

Conclusion

Explaining human causal induction requires supplementing the formal methods developed in computer science with the causal domain knowledge that people possess. We have shown that using physical theories to inform rational statistical inference makes it possible to explain how people infer hidden causal structure from such limited data. We anticipate that the same framework, using appropriately modified causal theories, can shed light on inferences about hidden causes in other domains.

Acknowledgments We thank T. Kushnir and L. Schulz for helpful discussions. TLG was supported by a Stanford Graduate Fellowship, JBT by the P.E.Newton chair.

References

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

Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology*. MIT Press, Cambridge, MA.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., and Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, 111:1–31.

Hagmayer, Y. and Waldmann, M. R. (2002). How temporal assumptions influence causal judgments. *Memory and Cognition*, 30:1128–1137.

Kushnir, T., Gopnik, A., Schulz, L., and Danks, D. (2003). Inferring hidden causes. In *Proceedings of the 25th Conference of the Cognitive Science Society*.

Luhmann, C. C. and Ahn, W.-K. (2003). Evaluating the causal role of unobserved variables. In *Proceedings of the 25th Conference of the Cognitive Science Society*.

Newsome, G. L. (2003). The debate between current versions of the covariation and mechanism approaches to causal inference. *Philosophical Psychology*, 16:87–107.

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

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

Rozenblit, L. R. and Keil, F. C. (2002). The missunderstood limits of folk science: an illusion of explanatory depth. *Cognitive Science*, 26:521–562.

Shultz, T. R. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development*, 47(Serial no. 194).

Spirtes, P., Glymour, C., and Schienes, R. (1993). *Causation prediction and search*. Springer-Verlag, NY.

[1] A full derivation of this result is available at http://www-psych.stanford.edu/~gruffydd/reports/nitrox.pdf