Modeling causal learning in children

Tom Griffiths
UC Berkeley

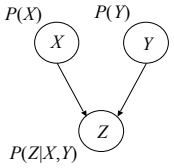
joint work with Josh Tenenbaum, Dave Sobel, and Alison Gopnik

Rational analysis of causal induction

- Computational problem: inferring causal relationships from observed data

  (Anderson, 1990; Anderson & Sheu, 1995; Griffiths &Tenenbaum, 2005; Steyvers et al., 2003; Waldmann & Martignon, 1998)
- Two questions:
  - how can we represent causal relationships?
  - how should we make the inference?

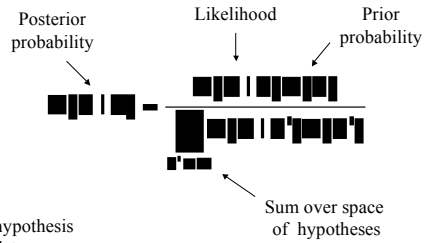
Causal graphical models



- Variables
- Structure
- Conditional probabilities

Defines probability distribution over variables (for both observation, and intervention)

Bayes' theorem



h: hypothesis
d: data

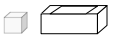
The puzzle

- How do children learn so much (rich causal structure) from so little (limited data)?

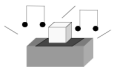
Blicket detector

(Dave Sobel, Alison Gopnik, and colleagues)

See this? It's a blicket machine. Blickets make it go.



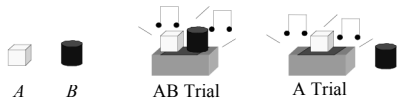
Let's put this one on the machine.



Oooh, it's a blicket!

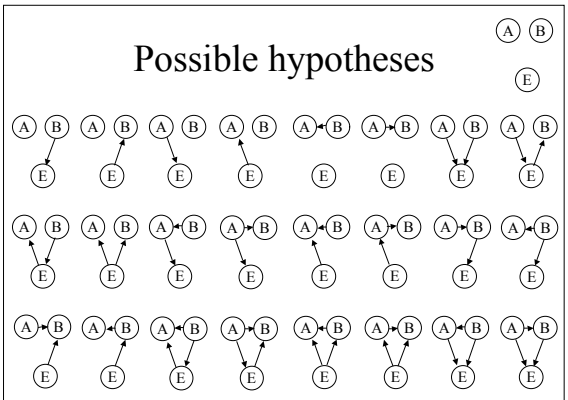
"Backwards blocking"

(Sobel, Tenenbaum & Gopnik, 2004)



- Two objects: A and B
- Trial 1: A B on detector – detector active
- Trial 2: A on detector – detector active
- 4-year-olds judge whether each object is a blicket
  - A: a blicket (100% say yes)
  - B: probably not a blicket (34% say yes)

Possible hypotheses



Bayesian inference

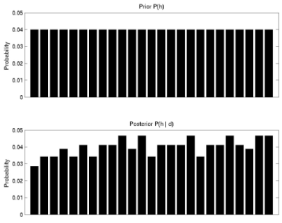
- Evaluating causal models in light of data:

- Inferring a particular causal relation:

Bayesian inference

With a uniform prior on hypotheses, and a simple "binomial" parameterization, integrating over parameters

(Cooper & Herskovits, 1992)



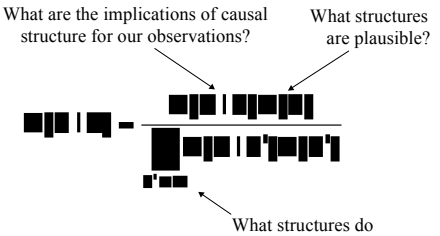
Probability of being a blicket

A	B
0.32	0.32
0.34	0.34

The puzzle

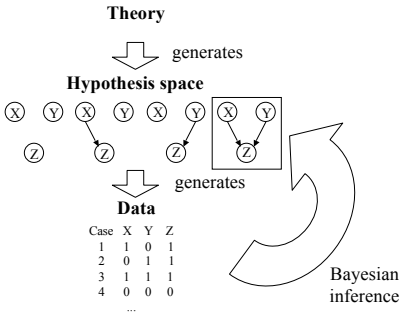
- How do children learn so much (rich causal structure) from so little (limited data)?
- Standard machine learning methods (and many psychological theories) do not exploit constraints from prior knowledge…

Constraints from prior knowledge

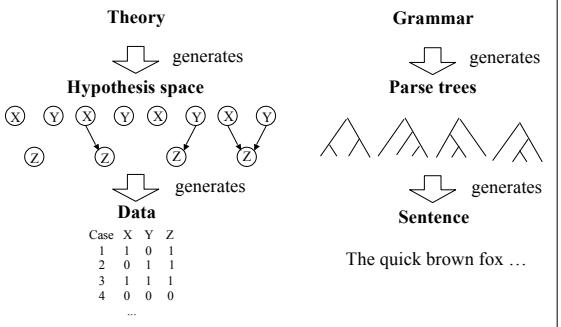


h: hypothesis
d: data

Theory-based causal induction



An analogy to language



The influence of prior knowledge

(e.g., Waldmann, 1996; Lagnado & Sloman, 2004)

- Prior knowledge produces expectations about:
  - types of entities
  - plausible relations
  - functional form
- This cannot be captured by graphical models

A theory consists of three interrelated components: a set of phenomena that are in its domain, the **causal laws** and other explanatory mechanisms in terms of which the phenomena are accounted for, and the **concepts** in terms of which the phenomena and explanatory apparatus are expressed. (Carey, 1985)

Theory-based causal induction

A causal theory is a *hypothesis space generator*

Component of theory:	**Generates:**
• Ontology	• Variables
• Plausible relations	• Structure
• Functional form	• Conditional probabilities

Hypotheses are evaluated by Bayesian inference

$$P(h|\text{data}) \propto P(\text{data}|h)\,P(h)$$

Theory

- **Ontology**
  - **Types:** Block, Detector, Trial
  - **Predicates:**
    Contact(Block, Detector, Trial)
    Active(Detector, Trial)

Theory

- **Ontology**
  - **Types:** Block, Detector, Trial
  - **Predicates:**
    Contact(Block, Detector, Trial)
    Active(Detector, Trial)

A = 1 if Contact(block A, detector, trial), else 0
B = 1 if Contact(block B, detector, trial), else 0
E = 1 if Active(detector, trial), else 0

Theory

- **Plausible relations**
  - For any Block b and Detector d, with prior probability q : For all trials t, Contact(b,d,t) → Active(d,t)

No hypotheses with E → B, E → A, A → B, etc.

$P(h_{00}) = (1-q)^2$ h_{00} : A B E

$P(h_{10}) = q(1-q)$ h_{10} : A B E

A → E = "A is a blicket"

$P(h_{01}) = (1-q)\,q$ h_{01} : A B E

$P(h_{11}) = q^2$ h_{11} : A B E

Theory

- **Functional form of causal relations**
  - Causes of Active(d,t) are independent mechanisms, with causal strengths w_b. A background cause has strength w_0. Assume a deterministic mechanism: $w_b = 1$, $w_0 = 0$.

	$P(h_{00}) = (1-q)^2$	$P(h_{01}) = (1-q)\,q$	$P(h_{10}) = q(1-q)$	$P(h_{11}) = q^2$
$P(E=1 \mid A=0, B=0)$:	0	0	0	0
$P(E=1 \mid A=1, B=0)$:	0	0	1	1
$P(E=1 \mid A=0, B=1)$:	0	1	0	1
$P(E=1 \mid A=1, B=1)$:	0	1	1	1

Theory

- **Ontology**
  - **Types:** Block, Detector, Trial
  - **Predicates:**
    Contact(Block, Detector, Trial)
    Active(Detector, Trial)
- **Plausible relations**
  - For any Block b and Detector d, with prior probability q : For all trials t, Contact(b,d,t) → Active(d,t)
- **Functional form of causal relations**
  - Causes of Active(d,t) are independent mechanisms, with causal strengths w_i. A background cause has strength w_0. Assume a deterministic mechanism: $w_b = 1$, $w_0 = 0$.

Bayesian inference

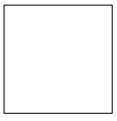
- Evaluating causal models in light of data:

- Inferring a particular causal relation:

Modeling backwards blocking

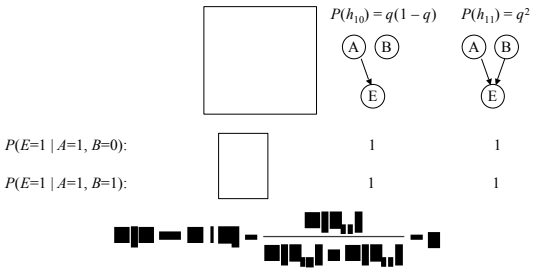
	$P(h_{00}) = (1-q)^2$	$P(h_{01}) = (1-q)\,q$	$P(h_{10}) = q(1-q)$	$P(h_{11}) = q^2$
$P(E=1 \mid A=0, B=0)$:	0	0	0	0
$P(E=1 \mid A=1, B=0)$:	0	0	1	1
$P(E=1 \mid A=0, B=1)$:	0	1	0	1
$P(E=1 \mid A=1, B=1)$:	0	1	1	1

Modeling backwards blocking



	$P(h_{01}) = (1-q)\,q$	$P(h_{10}) = q(1-q)$	$P(h_{11}) = q^2$
$P(E=1 \mid A=1, B=1)$:	1	1	1

Modeling backwards blocking

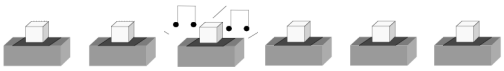


Theory

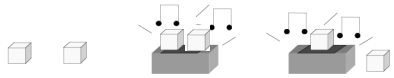
- **Ontology**
  - **Types:** Block, Detector, Trial
  - **Predicates:**
    Contact(Block, Detector, Trial)
    Active(Detector, Trial)
- **Plausible relations**
  - For any Block b and Detector d, with prior probability q :
    For all trials t, Contact(b,d,t) → Active(d,t)
- **Functional form of causal relations**
  - Causes of Active(d,t) are independent mechanisms, with causal strengths w_i. A background cause has strength w_0. Assume a deterministic mechanism: $w_i = 1$, $w_0 = 0$.

Manipulating plausibility

I. Pre-training phase: Establish baserate for blickets (q)



II. Backwards blocking phase:



After each trial, adults judge the probability that each object is a blicket.

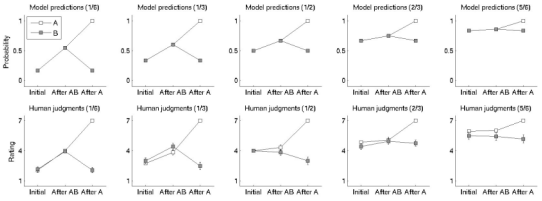
Manipulating plausibility

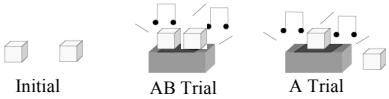
- Expose to different base-rates
  - q = 1/6, 1/3, 1/2, 2/3, 5/6
- Test with backwards blocking
- Model makes two qualitative predictions:
  - evaluation of both A and B as blickets will increase with baserate
  - evaluation of B will increase after AB Trial, then return to baserate after A Trial

(Tenenbaum, Sobel, Griffiths, & Gopnik, submitted)

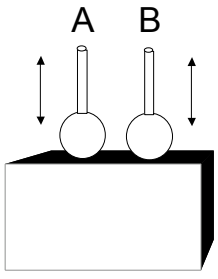
Manipulating plausibility

(n = 12 per condition)





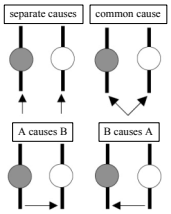
The stick-ball machine



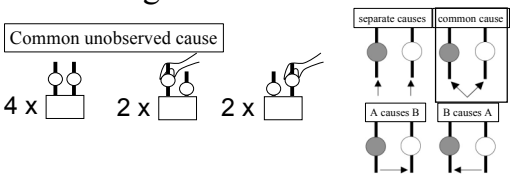
(Kushnir, Schulz, Gopnik, & Danks, 2003)

Inferring hidden causal structure

- Can people accurately infer hidden causal structure from small amounts of data?
- Kushnir et al. (2003): four kinds of structure

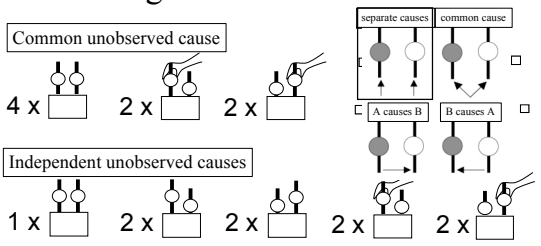


Inferring hidden causal structure



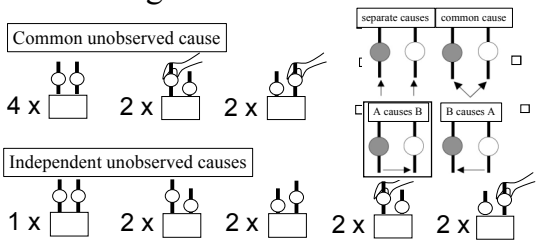
(Kushnir, Schulz, Gopnik, & Danks, 2003)

Inferring hidden causal structure

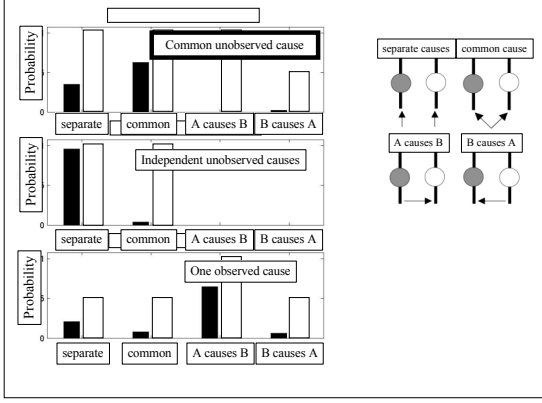


(Kushnir, Schulz, Gopnik, & Danks, 2003)

Inferring hidden causal structure

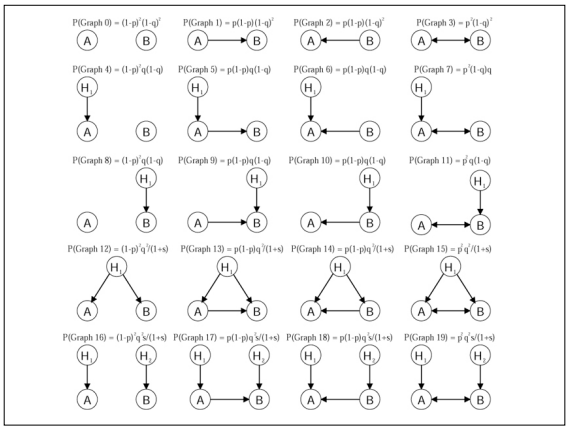


(Kushnir, Schulz, Gopnik, & Danks, 2003)

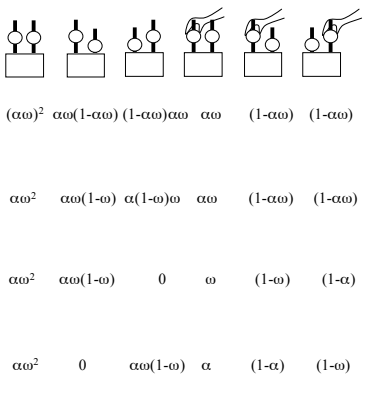


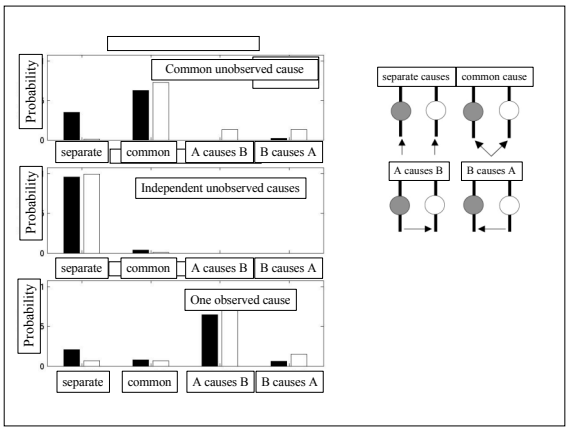
Theory

- **Ontology**
  - **Types:** Ball, HiddenCause, Trial
  - **Predicates:** Moves(Ball, Trial), Active(HiddenCause, Trial)
- **Plausible relations**
  - For any Ball a and Ball b ($a \neq b$), with prior probability p: For all Trials t, Moves(a,t) → Moves(b,t)
  - For some HiddenCause h and Ball b, with prior probability q: For all Trials t, Active(h,t) → Moves(b,t)
- **Functional form of causal relations**
  - Causes result in Moves(b,t) with probability ω. Otherwise, Moves(b,t) occurs with probability 0.
  - Active(h,t) occurs with probability α.

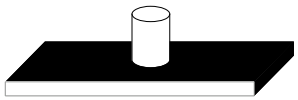


Hypotheses

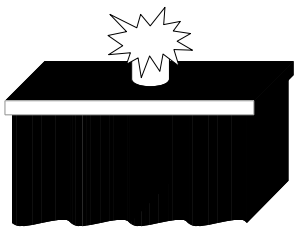


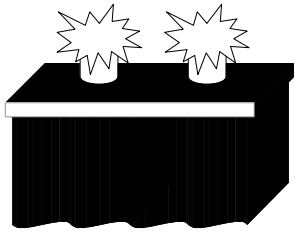


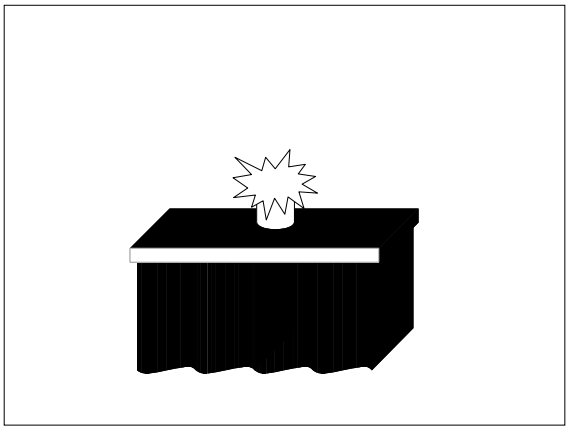
Nitro X



- More extreme test of ability to infer hidden causes
  - single datapoint
  - no mention of hidden cause in instructions
- More sophisticated physical theory
- Importance of *statistical* inference

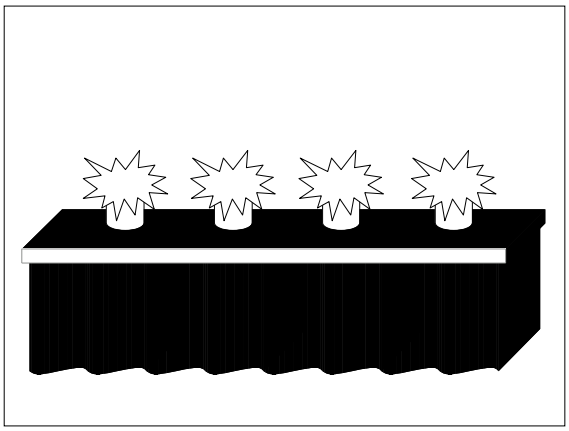


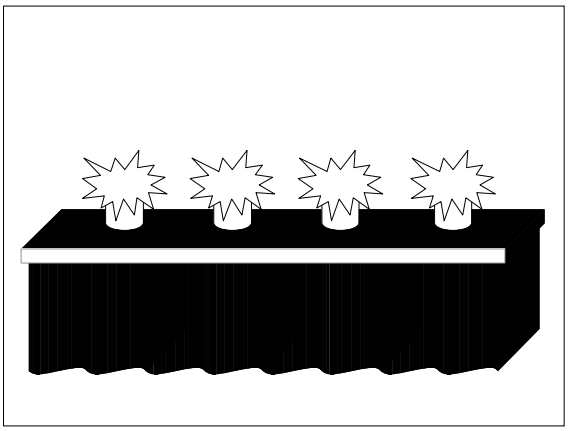


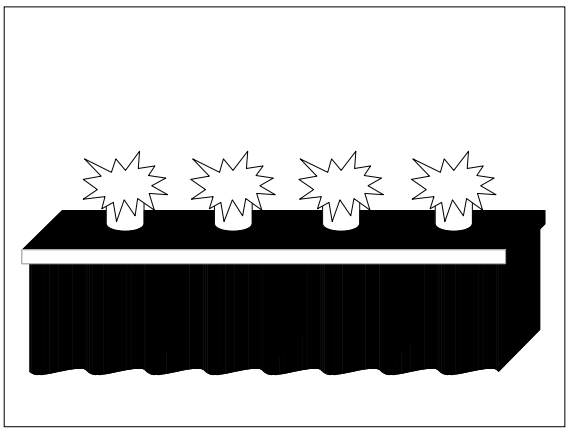


Test trials

- Show explosions involving multiple cans
  - allows inferences about causal structure
- For each trial, choose one of:
  - chain reaction
  - spontaneous explosions
  - other

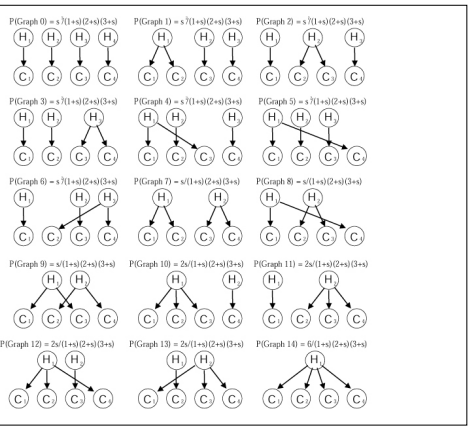






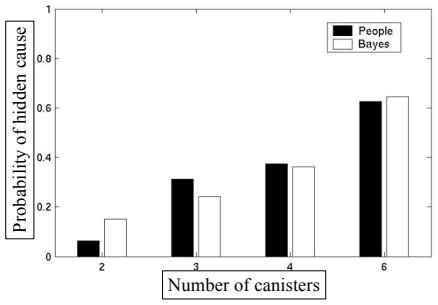
Theory

- **Ontology**
  - **Types:** Can, HiddenCause
  - **Predicates:**
    ExplosionTime(Can), ActivationTime(HiddenCause)
- **Plausible relations**
  - For any Can *y* and Can *x*, with prior probability 1: ExplosionTime(*y*) → ExplosionTime(*x*)
  - For some HiddenCause *c* and Can *x*, with prior probability 1: ActivationTime(*c*) → ExplosionTime(*x*)
- **Functional form of causal relations**
  - Explosion at ActivationTime(*c*), and after appropriate delay from ExplosionTime(*y*) with probability set by ω. Otherwise explosions occur with probability 0.
  - Low probability of hidden causes activating.



Testing the model predictions

- 16 participants in each of 4 conditions, varying number of canisters exploding simultaneously
- For each trial, chose one of:
  - The first can exploded spontaneously. That explosion caused the other cans to explode, in a chain reaction
  - Each can exploded spontaneously, all on its own. There was no causal connection between them
  - Neither of the above is a plausible explanation. Please write a plausible alternative here



$\chi^2(3) = 11.36$, $p < .01$

Conclusion

Children (and adults) can learn from small amounts of data by exploiting strong prior knowledge

rational statistical inference

$$\text{induction} = f(\text{knowledge}, \text{data})$$

intuitive theory
inductive bias

Rational analysis provides a way to determine this knowledge, identifying human inductive biases