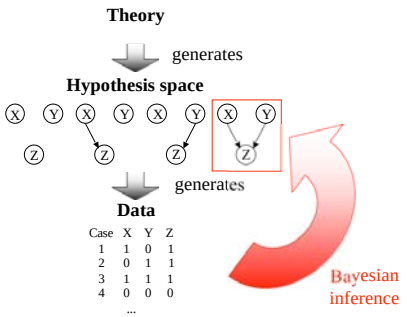
The development of causal theories

Tom Griffiths
UC Berkeley

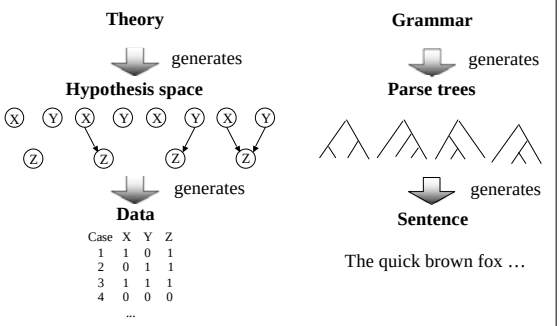
The puzzle

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

Theory-based causal induction



An analogy to language



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!

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:

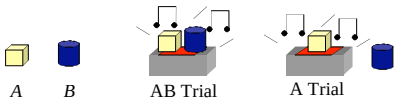
$$P(h_i \mid d) = \frac{P(d \mid h_i)P(h_i)}{\sum_j P(d \mid h_j)P(h_j)}$$

- Inferring a particular causal relation:

$$P(A \rightarrow E \mid d) = \sum_{h_j \in H} P(A \rightarrow E \mid h_j)P(h_j \mid d)$$

"Backwards blocking"

(Sobel, Tenenbaum & Gopnik, 2004)

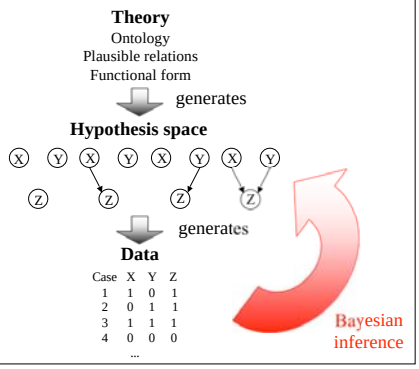


– Two objects: A and B
– Trial 1: A B on detector – detector active
– Trial 2: A on detector – detector active
– 4-year-olds judge whether each object is a blicket
  • A: a blicket (100% say yes)
  • B: probably not a blicket (34% say yes)

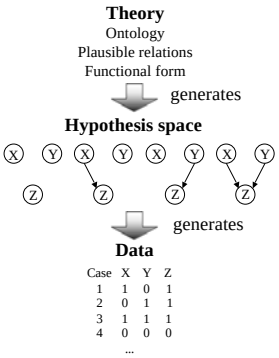
The new puzzle

- How do people learn so much (causal theories) from so little (limited data)?

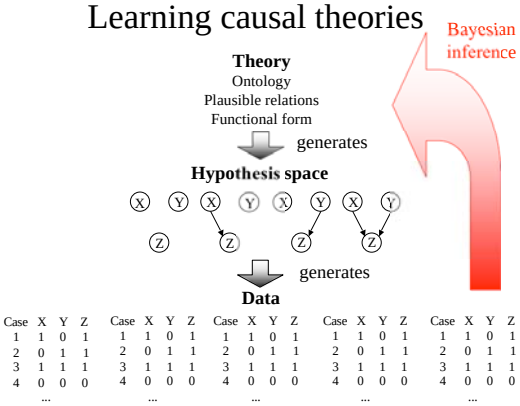
Learning causal theories



Learning causal theories



Learning causal theories



Pushing the grammar analogy

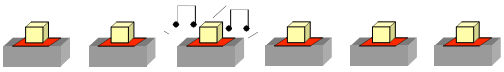
- The ways of learning the parts of causal theories will be similar to methods for learning grammars
  - learning ontologies and nonterminals
  - learning plausible relations and production rules
  - learning plausibilities and parameters

Theory

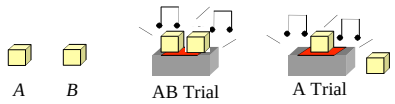
- **Ontology**
  - **Types:** Block, Detector, Trial
  - **Predicates:**
    Contact(Block, Detector, Trial)
    Active(Detector, Trial)
- **Plausible relations**
  - For any Block b and Detector d, with prior probability q :
    For all trials t, Contact(b,d,t) → Active(d,t)
- **Functional form of causal 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)



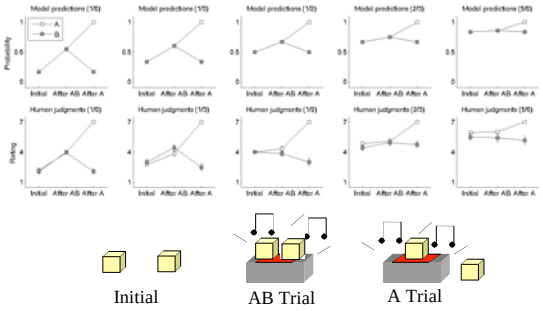
II. Backwards blocking phase:



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

Manipulating plausibility

(n = 12 per condition)



Results with children

- Tested 32 four-year-olds (mean age 53 months)
- Instead of rating, yes or no response
- Two conditions
  - blickets are rare, 2/12 in familiarization phase
  - blickets are common, 10/12 in familiarization phase
- Significant difference in one cause *B* responses
  - rare: 25% say yes
  - common: 81% say yes

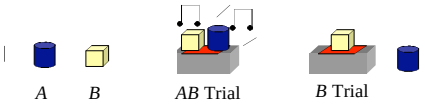
(Sobel, Tenenbaum, & Gopnik, 2004)

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: $w_b = 1$, $w_0 = 0$.

"One cause"

(Gopnik, Sobel, Schulz, & Glymour, 2001)

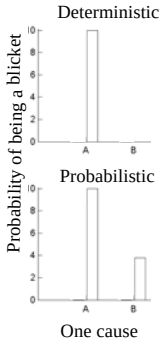


– Two objects: A and B
– Trial 1: A B on detector – detector active
– Trial 2: B on detector – detector inactive
– 4-year-olds judge whether each object is a blicket
  - *A*: a blicket (100% say yes)
  - *B*: almost certainly not a blicket (16% say yes)

A probabilistic mechanism?

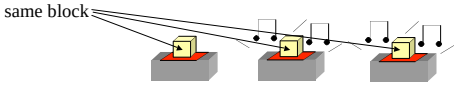
- Children in Gopnik et al. (2001) who said that *B* was a blicket had seen evidence that the detector was probabilistic
  – one block activated detector 5/6 times
- Replace the deterministic "activation law"…
  – activate with $p = 1-\varepsilon$ if a blicket is on the detector
  – never activate otherwise

Deterministic vs. probabilistic

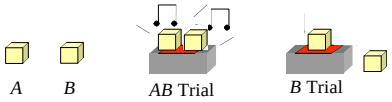


Manipulating functional form

I. Familiarization phase: Establish nature of mechanism



II. Test phase: one cause



At end of the test phase, adults judge the probability that each object is a blicket

Manipulating functional form

- Expose to different kinds of functional form
  – deterministic: detector always activates
  – probabilistic: detector activates with $p = 1-\varepsilon$
- Test with "one cause" trials
- Model makes two qualitative predictions:
  – people will infer functional form
  – evaluation of *B* as a blicket will increase with the probabilistic mechanism

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

Learning causal theories

- Apply Bayes' rule as before:

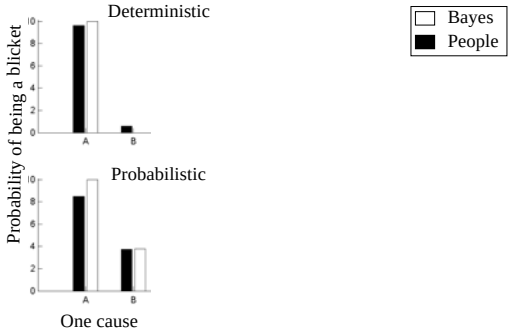
$$P(T_i \mid d) = \frac{P(d \mid T_i)P(T_i)}{\sum_j P(d \mid T_j)P(T_j)}$$

- Sum over causal structures (h_j) to get $P(d|T)$

$$P(d \mid T) = \sum_{h_j \in H_T} P(d \mid h_j)P(h_j \mid T)$$

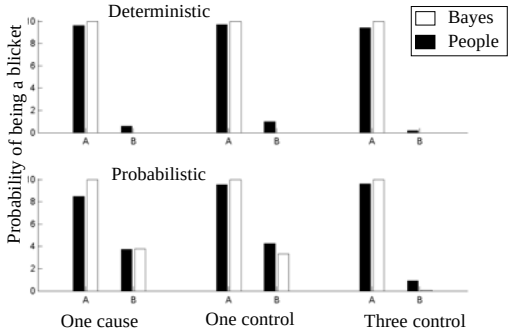
Manipulating functional form

(n = 12 undergraduates per condition)



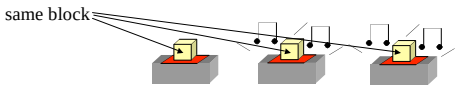
Manipulating functional form

(n = 12 undergraduates per condition)

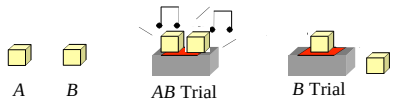


Acquiring mechanism knowledge

I. Familiarization phase: Establish functional form



II. Test phase: one cause



At end of the test phase, adults judge the probability that each object is a blicket

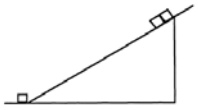
Results with children

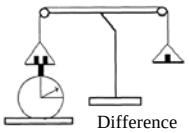
- Tested 24 four-year-olds (mean age 54 months)
- Instead of rating, yes or no response
- Significant difference in one cause *B* responses
  - deterministic: 8% say yes
  - probabilistic: 79% say yes
- No significant difference in one control trials
  - deterministic: 4% say yes
  - probabilistic: 21% say yes

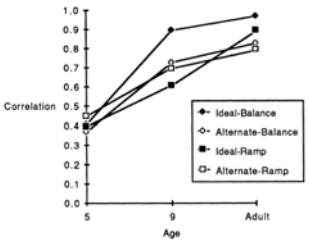
Summary

- Using causal systems like the blicket detector, we can teach people new parts of causal theories
  - plausibility of causal relationships
  - functional form of those relationships
- It is possible for one observation to produce a radical change in the causal theories maintained

Development of functional form





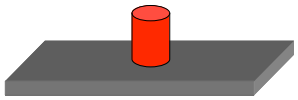


(Zelazo & Shultz, 1989)

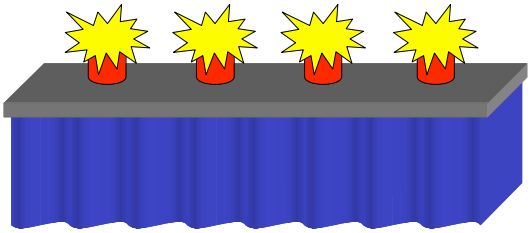
Summary

- Using causal systems like the blicket detector, we can teach people new parts of causal theories
  - plausibility of causal relationships
  - functional form of those relationships
- It is possible for one observation to produce a radical change in the causal theories maintained
- But what about more complex causal systems?
  - form of forces?
  - parameters of forces?
  - new forces?

Parameter estimation with Nitro X



For known causal forces, how do we estimate the constants that are relevant to the force?



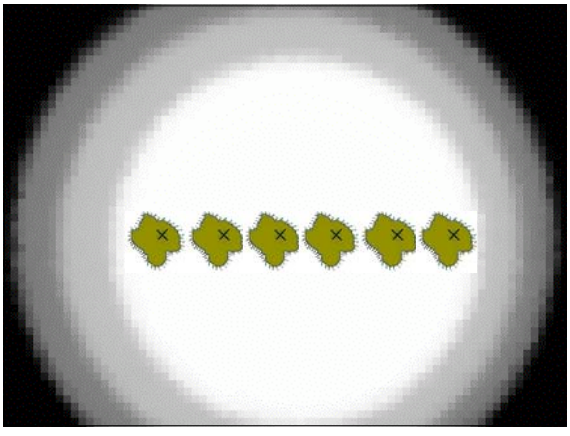
Theory

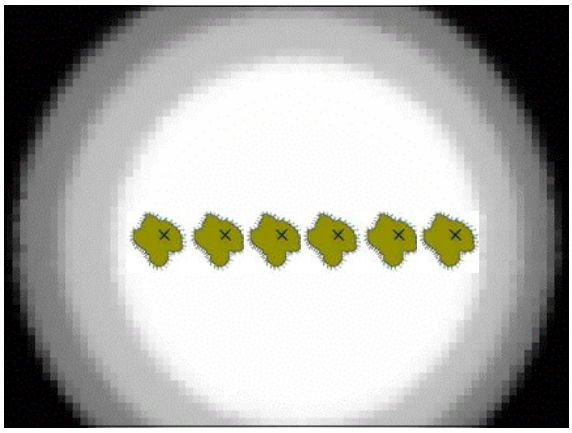
- **Ontology**
  - **Types:** Can, HiddenCause
  - **Predicates:**
    ExplosionTime(Can), ActivationTime(HiddenCause)
- **Plausible relations**
  - For any Can *y* and Can *x*, with prior probability 1: ExplosionTime(*y*) → ExplosionTime(*x*)
  - For some HiddenCause *c* and Can *x*, with prior probability 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.

New forces

How do people discover new kinds of causal relationships?

(A great deal of what we do in science)





Learning causal theories

- T_1: bacteria die at random
- T_2: bacteria die at random, or in waves

$$P(\text{wave}|T_2) > P(\text{wave}|T_1)$$

- Having inferred the existence of a new force, need to find a mechanism...

Shallow theories

- To learn and reason about causality, we need
  - functional form of causal relationship
  - knowledge that mechanisms exist
- We can figure out the mechanisms once we know that we need them…
- So we can get away with shallow theories
  - illusion of explanatory depth (Rozenblit & Keil. 2002)

Conclusion

- From a formal perspective, learning causal theories is just a matter of pushing Bayes up the hierarchy
- But… understanding the development of causal theories requires understanding the kinds of knowledge that constitute those theories
  - minimally: ontology, plausibility, functional form
- Sensitivity to "coincidences" is key, as the clue to search for a plausible mechanism…

Challenges

- What are hypothesis spaces of causal theories?
- Can we define theory generators, in the same ay that theories act as hypothesis generators?
- Constraints on learning are still going to be important, but…
  - hopefully less strong ("blessing of abstraction")
  - more plausibly innate