Approaches to structure learning

- Constraint-based learning (Pearl, Glymour, Gopnik):
  - Assume structure is unknown, no knowledge of parameterization or parameters
- Bayesian learning (Heckerman, Friedman/Koller):
  - Assume structure is unknown, arbitrary parameterization.
- Theory-based Bayesian inference (T & G):
  - Assume structure is partially unknown, parameterization is known but parameters may not be. *Prior knowledge about structure and parameterization depends on domain theories (derived from ontology and mechanisms).*

Advantages/Disadvantages of the constraint-based approach

- Deductive
- Domain-general
- No essential role for domain knowledge:
  - Knowledge of possible causal structures not needed.
  - Knowledge of possible causal mechanisms not used.
- Requires large sample sizes to make reliable inferences.

The Blicket detector

Image removed due to copyright considerations. Please see:
Gopnick, A., and D. M. Sobel. “Detecting Blickets: How Young Children use Information about Novel Causal Powers in Categorization and Induction.” *Child Development* 71 (2000): 1205-1222.

Image removed due to copyright considerations. Please see:
Gopnick, A., and D. M. Sobel. “Detecting Blickets: How Young Children use Information about Novel Causal Powers in Categorization and Induction.” *Child Development* 71 (2000): 1205-1222.

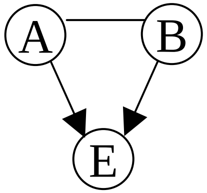
The Blicket detector

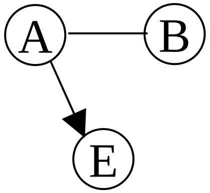
- Can we explain these inferences using constraint-based learning?
- What other explanations can we come up with?

Constraint-based model

- Data:
  - d_0: A=0, B=0, E=0
  - d_1: A=1, B=1, E=1
  - d_2: A=1, B=0, E=1
- Constraints:
  - *A*, *B* not independent
  - *A*, *E* not independent
  - *B*, *E* not independent
  - *B*, *E* independent conditional on the presence of *A*
  - *A*, *E* not independent conditional on the absence of *B*
  - Unknown whether *B*, *E* independent conditional on the absence of *A*.
- Graph structures consistent with constraints:

Image removed due to copyright considerations. Please see: Gopnick, A., and D. M. Sobel. "Detecting Blickets: How Young Children use Information about Novel Causal Powers in Categorization and Induction." *Child Development* 71 (2000): 1205-1222.





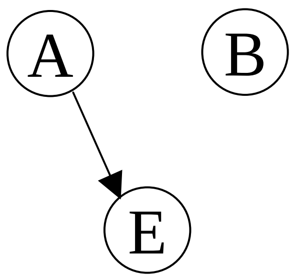
NOTE: Also have A, B independent conditional on the presence of E. Does that eliminate the hypothesis that B is a blicket?

Constraint-based inference

- Data:
  - d_1: A=1, B=1, E=1
  - d_2: A=1, B=0, E=1
  - d_0: A=0, B=0, E=0

Imagine sample sizes multiplied by 100….
(Gopnik, Glymour et al., 2002)

- Conditional independence constraints:
  - B, E independent conditional on A
  - B, A independent conditional on E
  - A, E correlated, unconditionally or conditional on B
- Inferred causal structure:



  - B is not a blicket.
  - A is a blicket.

Why not use constraint-based methods + fictional sample sizes?

- No degrees of confidence.
- No principled interaction between data and prior knowledge.
- Reliability becomes questionable.
  - “The prospect of being able to do psychological research without recruiting more than 3 subjects is so attractive that we know there must be a catch in it.”

A deductive inference?

- Causal law: detector activates if and only if one or more objects on top of it are blickets.
- Premises:
  - Trial 1: *A* *B* on detector – detector active
  - Trial 2: *A* on detector – detector active
- Conclusions deduced from premises and causal law:
  - *A*: a blicket
  - *B*: can't tell (Occam's razor → not a blicket?)

What kind of Occam's razor?

- Classical all-or-none form:
  - "Causes should not be multiplied without necessity."
- Constraint-based: faithfulness
- Bayesian: probability

For next time

- Come up with slides on Theory-based Bayesian causal inference.
- Combine current teaching slides, which emphasize Bayes versus constraint-based, with Leuven slides, which emphasize a systematic development of the theory.
- Incorporate (if time) cross-domains, plus AB-AC.

Approaches to structure learning

- Constraint-based learning (Pearl, Glymour, Gopnik):
  - Assume structure is unknown, no knowledge of parameterization or parameters

- Bayesian learning (Heckerman, Friedman/Koller):
  - Assume structure is unknown, arbitrary parameterization.

- Theory-based Bayesian inference (T & G):
  - Assume structure is partially unknown, parameterization is known but parameters may not be. *Prior knowledge about structure and parameterization depends on domain theories (derived from ontology and mechanisms).*

For next year

- Include deductive causal reasoning as one of the methods. It goes back a long time….

Critical differences between Bayesian and Constraint-based learning

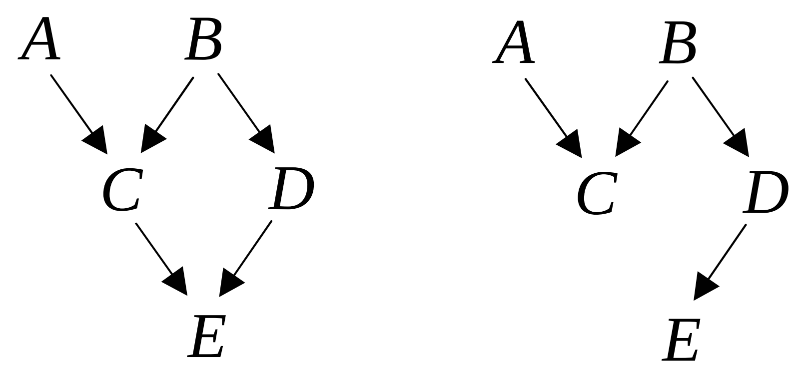
- Basis for inferences:
  - Constraint-based inference based on just qualitative independence constraints.
  - Bayesian inference based on full probabilistic models (generated by domain theory).
- Nature of inferences:
  - Constraint-based inferences are deductive.
  - Bayesian inferences are probabilistic.

Bayesian causal inference

Data X

$$x_1 = \langle A=1, B=1, C=1, D=1, E=1 \rangle$$
$$x_2 = \langle A=1, B=0, C=1, D=0, E=1 \rangle$$
$$x_3 = \langle A=0, B=1, C=0, D=1, E=0 \rangle$$
$$x_4 = \langle A=1, B=0, E=0 \rangle$$
$$x_5 = \langle C=1, E=1 \rangle$$

Causal hypotheses h



Bayes: $P(h \mid X) \propto P(X \mid h)\, P(h)$

Why be Bayesian?

- Explain how people can *reliably* acquire *true* causal beliefs given very limited data:
  - Prior causal knowledge: Domain theory
  - Causal inference procedure: Bayes

- Understand how symbolic domain theory interacts with rational statistical inference:
  - Theory generates the hypothesis space of candidate causal structures.

Role of domain theory

- Determines prior over models, $P(h)$
  - Causally relevant attributes of objects and relations between objects: variables
  - Viable causal relations: edges

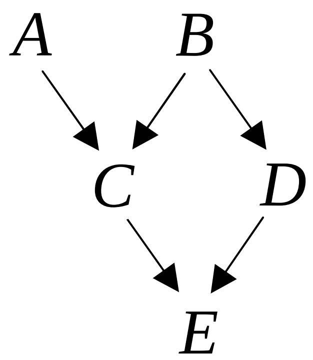
- Determines likelihood function for each model, $P(X|h)$, via (perhaps abstract or "light") mechanism knowledge:
  - How each effect depends functionally on its causes: $V \Leftarrow f_\theta(\text{parents}[V]) \longrightarrow P(V \mid \text{parents}[V])$

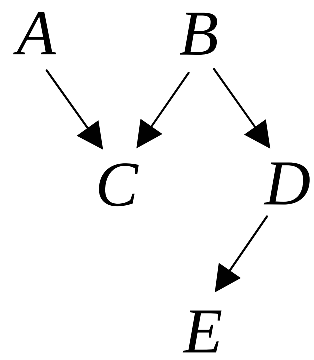
Bayesian causal inference

Data X

$$x_1 = \langle A=1, B=1, C=1, D=1, E=1 \rangle$$
$$x_2 = \langle A=1, B=0, C=1, D=0, E=1 \rangle$$
$$x_3 = \langle A=0, B=1, C=0, D=1, E=0 \rangle$$
$$x_4 = \langle A=1, B=0, E=0 \rangle$$
$$x_5 = \langle C=1, E=1 \rangle$$

Causal hypotheses h





Bayes: $P(h \mid X) \propto P(X \mid h)\, P(h)$

$$P(A, B, C, D, E \mid \text{causal model}) = \prod_{V \in \{A,B,C,D,E\}} P(V \mid \text{parents}[V])$$

(Bottom-up) Bayesian causal learning in AI

- Typical goal is data mining, with no strong domain theory.
  - Uninformative prior over models $P(h)$
  - Arbitrary parameterization (because no knowledge of mechanism), with no strong expectations of likelihoods $P(X|h)$.
- Results not that different from constraint-based approaches, other than more precise probabilistic representation of uncertainty.

“Backwards blocking”

(Sobel, Tenenbaum & Gopnik, 2004)

Image removed due to copyright considerations. Please see:
Gopnick, A., and D. M. Sobel. “Detecting Blickets: How Young Children use Information about Novel Causal Powers in Categorization and Induction.” *Child Development* 71 (2000): 1205-1222.

- 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% of judgments)
  - *B*: probably not a blicket (66% of judgments)

Theory

- **Ontology**
  - **Types:** Block, Detector, Trial
  - **Predicates:**

    Contact(Block, Detector, Trial)

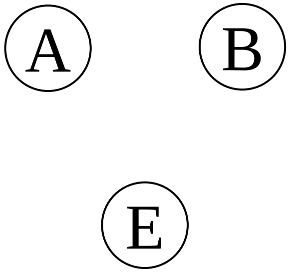
    Active(Detector, Trial)
- **Constraints on causal relations**
  - For any Block *b* and Detector *d*, with probability *q* : Cause(Contact(*b*,*d*,*t*), Active(*d*,*t*))
- **Functional form of causal relations**
  - Causes of Active(*d*,*t*) are independent mechanisms, with causal strengths w_i. A background cause has strength w_0. Assume a near-deterministic mechanism: $w_i \sim 1$, $w_0 \sim 0$.

Theory

- **Ontology**
  - **Types:** Block, Detector, Trial
  - **Predicates:**

    Contact(Block, Detector, Trial)

    Active(Detector, Trial)

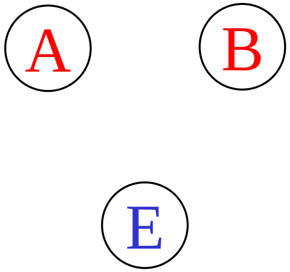


Theory

- **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

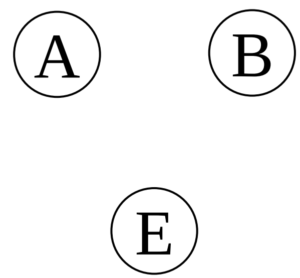
Theory

- **Constraints on causal relations**
  - For any Block *b* and Detector *d*, with probability *q* : Cause(Contact(*b*,*d*,*t*), Active(*d*,*t*))

No hypotheses with
$E \longrightarrow B$, $E \longrightarrow A$,
$A \longrightarrow B$, etc.

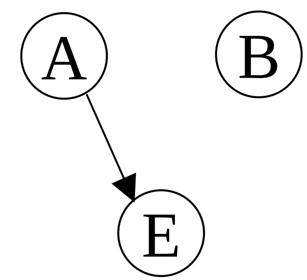
$P(h_{00}) = (1 - q)^2$

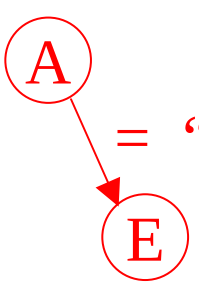
h_{00} :



$P(h_{10}) = q(1 - q)$

h_{10} :

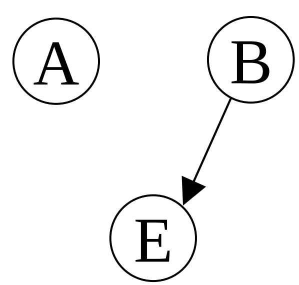




= "A is a blicket"

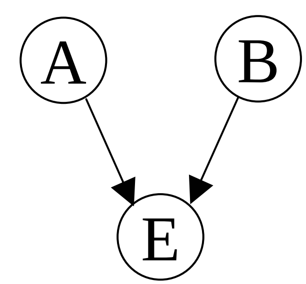
$P(h_{01}) = (1 - q)\, q$

h_{01} :



$P(h_{11}) = q^2$

h_{11} :



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 near-deterministic mechanism: $w_b \sim 1$, $w_0 \sim 0$.

$P(h_{00}) = (1 - q)^2$ $\quad$ $P(h_{01}) = (1 - q)\, q$ $\quad$ $P(h_{10}) = q(1 - q)$ $\quad$ $P(h_{11}) = q^2$

$P(E=1 \mid A=0, B=0)$:	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

"Activation law": E=1 if and only if A=1 or B=1.

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 near-deterministic mechanism: $w_b \sim 1$, $w_0 \sim 0$.

$P(h_{00}) = (1 - q)^2$ $P(h_{01}) = (1 - q)\, q$ $P(h_{10}) = q(1 - q)$ $P(h_{11}) = q^2$

$P(E{=}1 \mid A{=}0, B{=}0)$:	w_0	w_0	w_0	w_0
$P(E{=}1 \mid A{=}1, B{=}0)$:	w_0	w_0	$w_b + (1 - w_b)\, w_0$	$w_b + (1 - w_b)\, w_0$
$P(E{=}1 \mid A{=}0, B{=}1)$:	w_0	$w_b + (1 - w_b)\, w_0$	w_0	$w_b + (1 - w_b)\, w_0$
$P(E{=}1 \mid A{=}1, B{=}1)$:	w_0	$w_b + (1 - w_b)\, w_0$	$w_b + (1 - w_b)\, w_0$	$1 - (1 - w_b)^2 (1 - w_o)$

“Noisy-OR law”

Bayesian inference

- Evaluating causal network hypotheses in light of data:

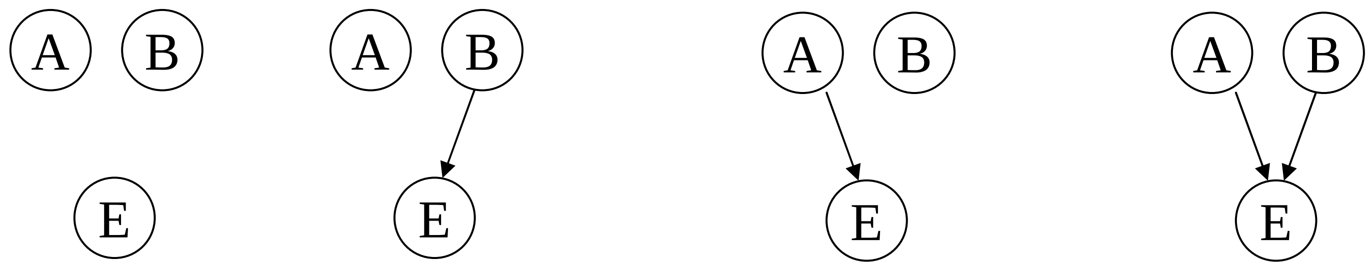
$$P(h_i \mid d) = \frac{P(d \mid h_i) P(h_i)}{\sum_{h_j \in H} P(d \mid h_j) P(h_j)}$$

- Inferring a particular causal relation:

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

Modeling backwards blocking

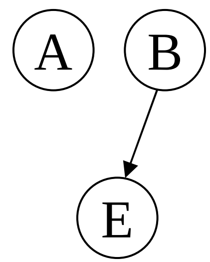
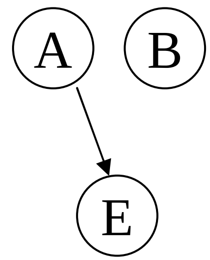
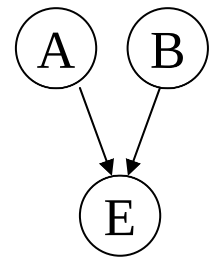
$P(h_{00}) = (1 - q)^2$ $P(h_{01}) = (1 - q)\, q$ $P(h_{10}) = q(1 - q)$ $P(h_{11}) = q^2$



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

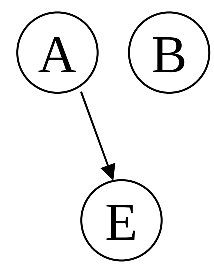
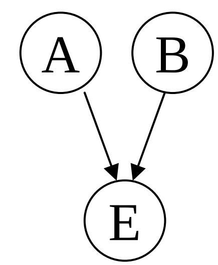
$$\frac{P(B \rightarrow E \mid d)}{P(B \qquad E \mid d)} = \frac{P(h_{01}) + P(h_{11})}{P(h_{00}) + P(h_{10})} = \frac{q}{1-q}$$

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

$$\frac{P(B \rightarrow E \mid d)}{P(B \qquad E \mid d)} = \frac{P(h_{01}) + P(h_{11})}{P(h_{10})} = \frac{1}{1-q}$$

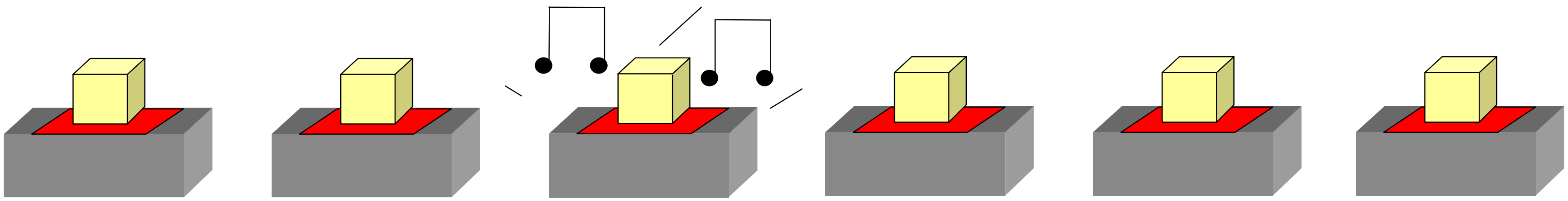
Modeling backwards blocking

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

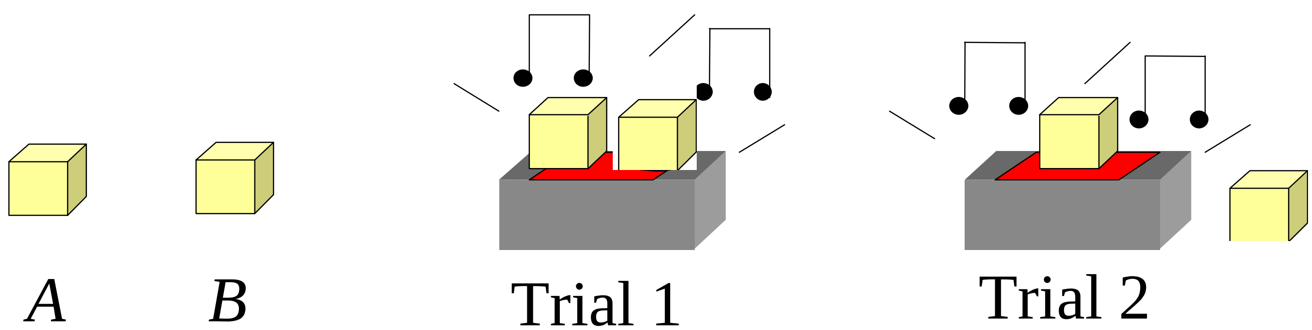
$$\frac{P(B \to E \mid d)}{P(B \qquad E \mid d)} = \frac{P(h_{11})}{P(h_{10})} = \frac{q}{1-q}$$

Manipulating the prior

I. Pre-training phase: Blickets are rare . . . .



II. Backwards blocking phase:



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

- “Rare” condition: First observe 12 objects on detector, of which 2 set it off.

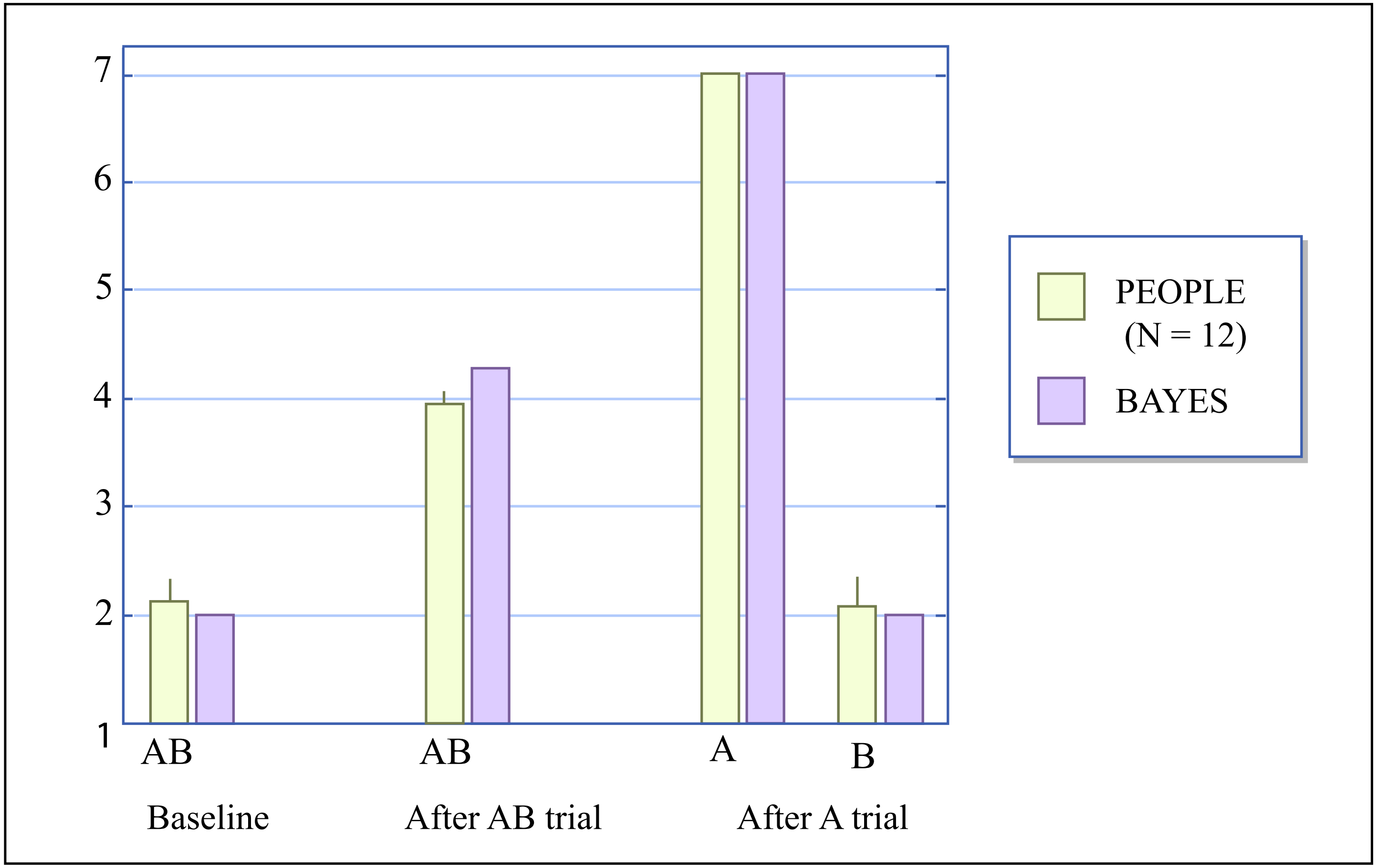


Figure by MIT OCW.

- "Common" condition: First observe 12 objects on detector, of which 10 set it off.

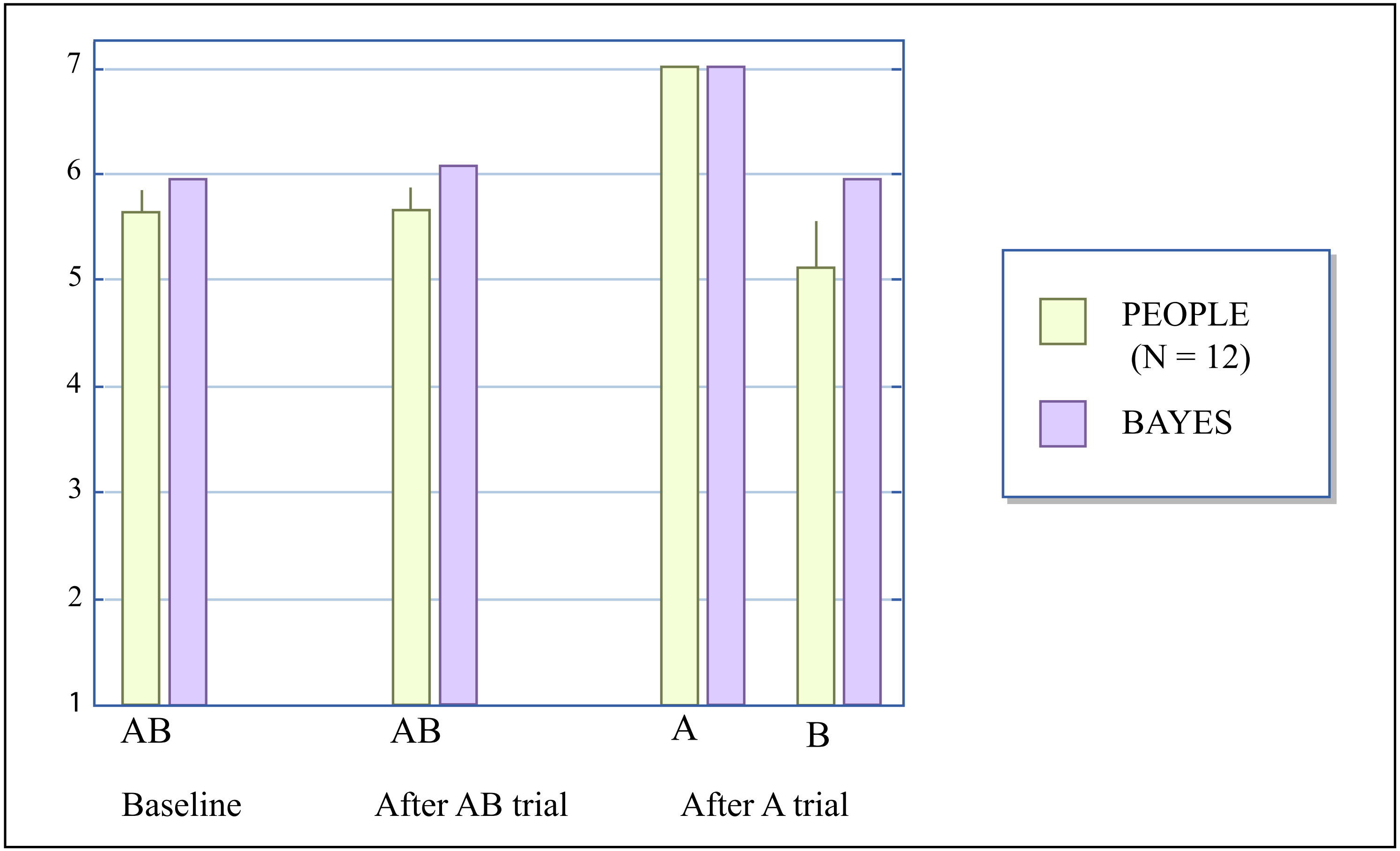


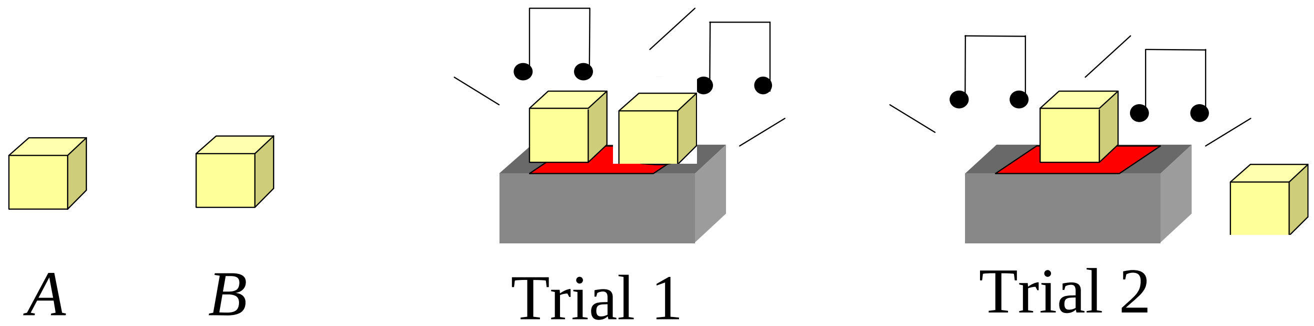
Figure by MIT OCW.

Manipulating the priors of 4-year-olds

(Sobel, Tenenbaum & Gopnik, 2004)

I. Pre-training phase: Blickets are rare.

II. Backwards blocking phase:



Rare condition:

A: 100% say “a blicket”

B: 25% say “a blicket”

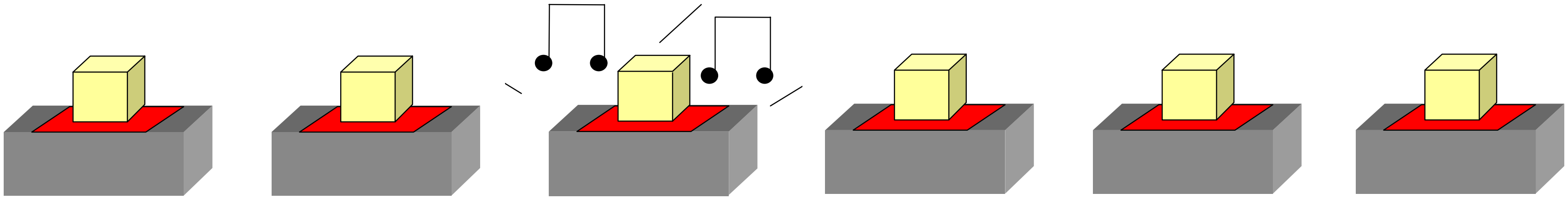
Common condition:

A: 100% say “a blicket”

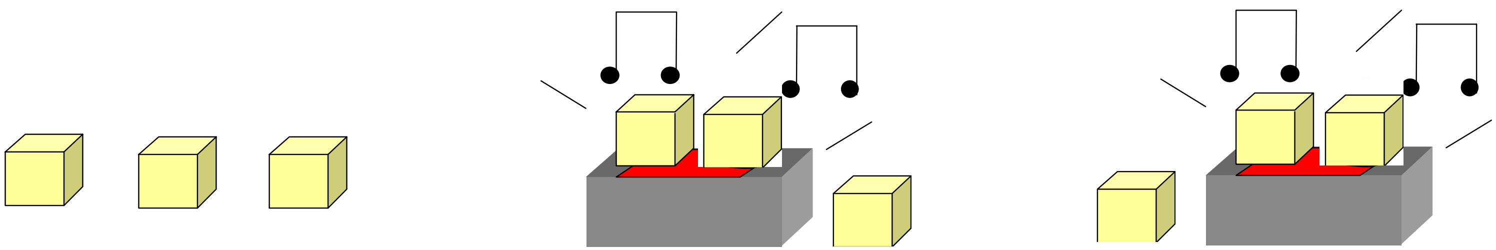
B: 81% say “a blicket”

Inferences from ambiguous data

I. Pre-training phase: Blickets are rare . . . .



II. Two trials: A B $\longrightarrow$ detector, B C $\longrightarrow$ detector

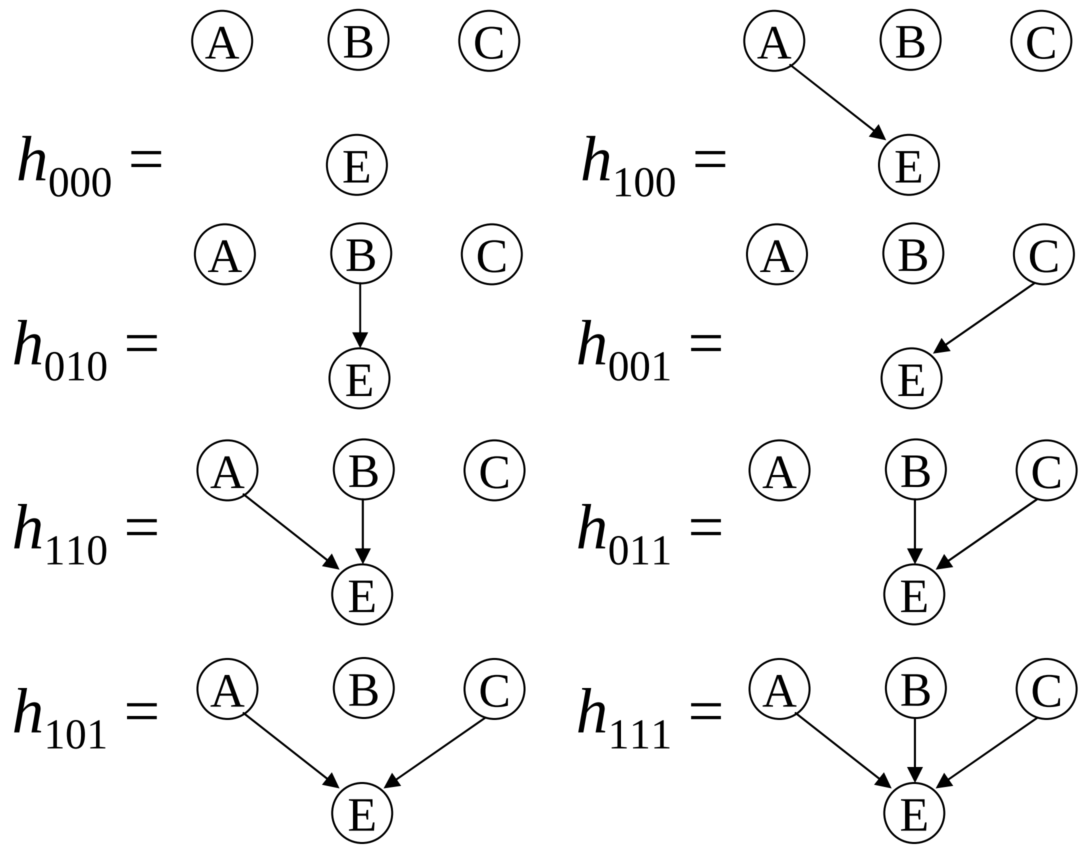


A   *B*   *C*   Trial 1   Trial 2

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

Same domain theory generates hypothesis space for 3 objects:

- Hypotheses:



- Likelihoods: $P(E{=}1 | A, B, C; h) = 1$ if $A = 1$ and $A \longrightarrow E$ exists,
  or $B = 1$ and $B \longrightarrow E$ exists,
  or $C = 1$ and $C \longrightarrow E$ exists,
  else 0.

- “Rare” condition: First observe 12 objects on detector, of which 2 set it off.

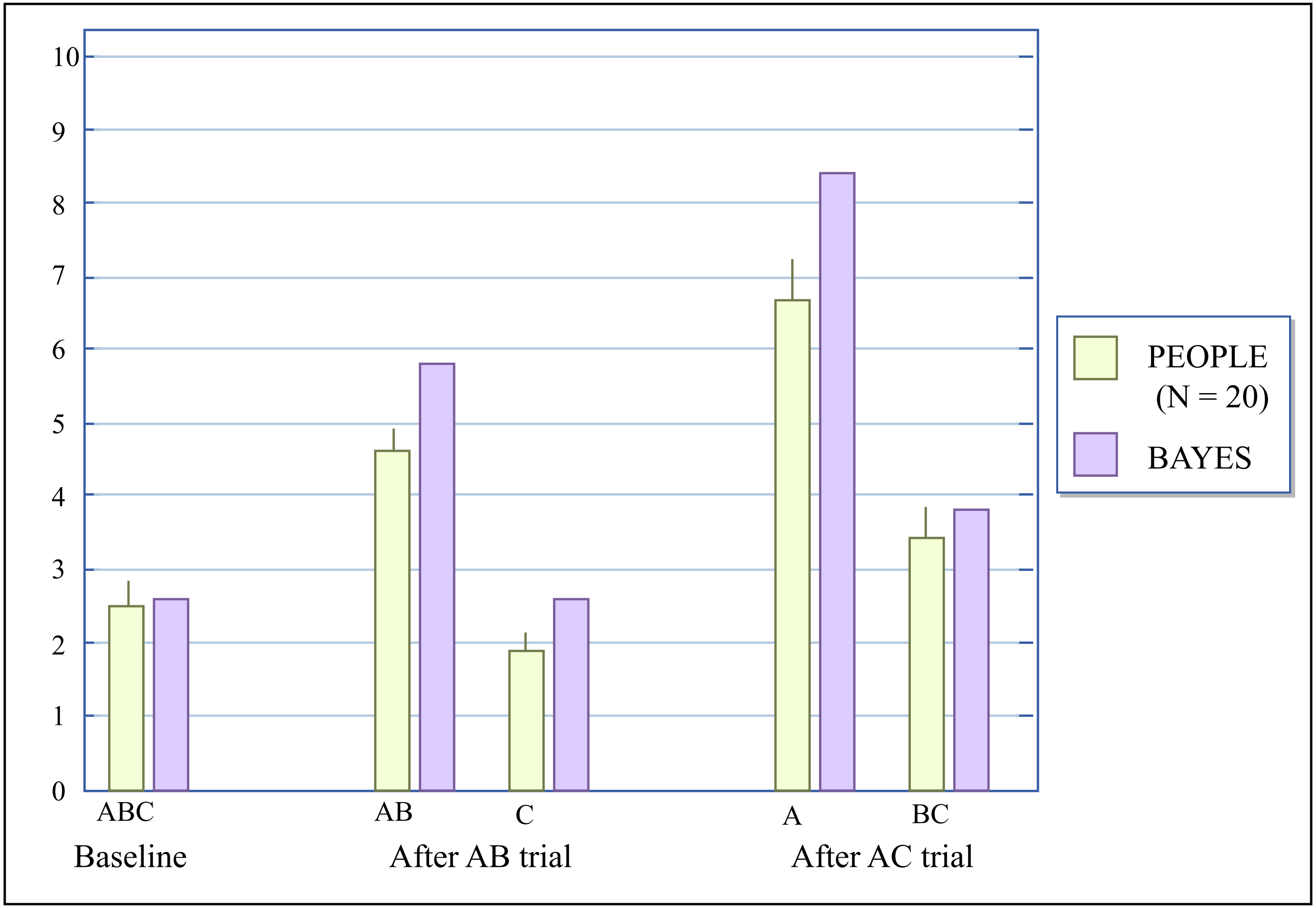
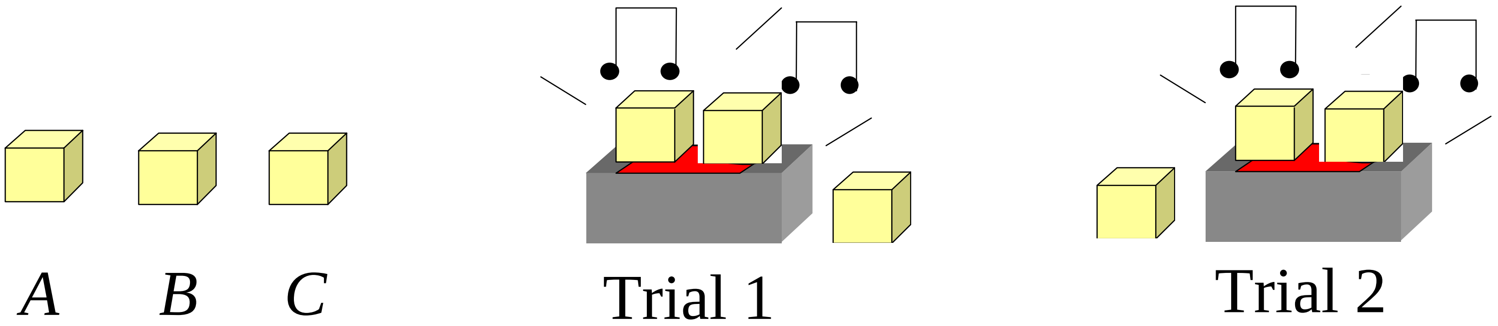


Figure by MIT OCW.

Ambiguous data with 4-year-olds

I. Pre-training phase: Blickets are rare.

II. Two trials: A B $\longrightarrow$ detector, B C $\longrightarrow$ detector



Final judgments:

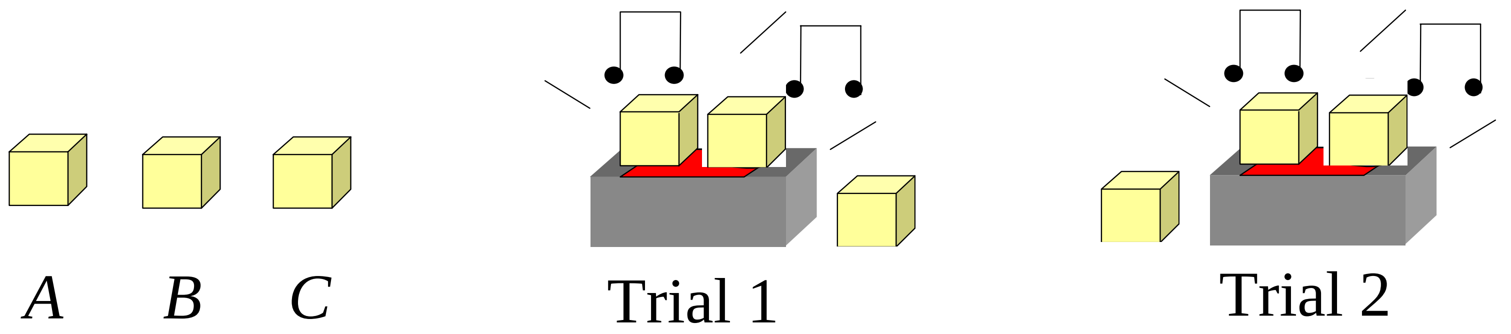
A: 87% say “a blicket”

B or *C*: 56% say “a blicket”

Ambiguous data with 4-year-olds

I. Pre-training phase: Blickets are rare.

II. Two trials: A B $\rightarrow$ detector, B C $\rightarrow$ detector



Final judgments:	Backwards blocking (rare)
A: 87% say “a blicket”	*A*: 100% say “a blicket”
B or *C*: 56% say “a blicket”	*B*: 25% say “a blicket”

The role of causal mechanism knowledge

- Is mechanism knowledge necessary?
  - Constraint-based learning using χ^2 tests of conditional independence.

- How important is the deterministic functional form of causal relations?
  - Bayes with “probabilistic independent generative causes” theory (i.e., noisy-OR parameterization with unknown strength parameters; c.f., Cheng’s causal power).

Bayes with correct theory:

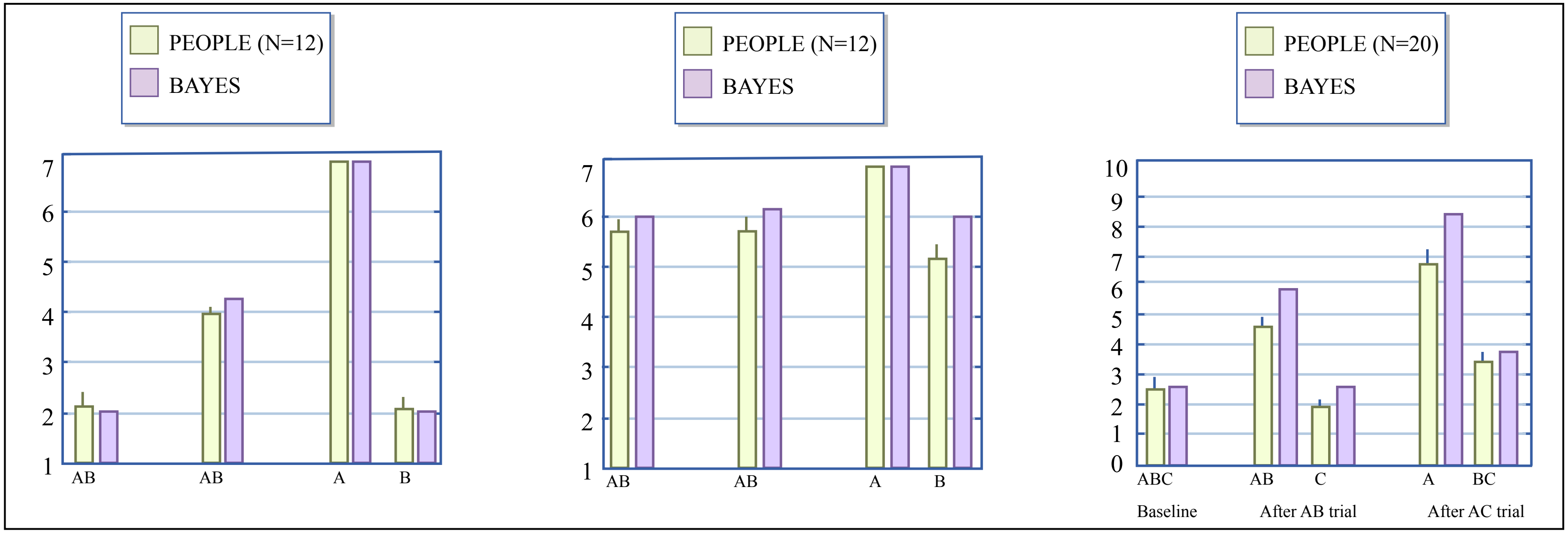


Figure by MIT OCW.

Independence test with fictional sample sizes:

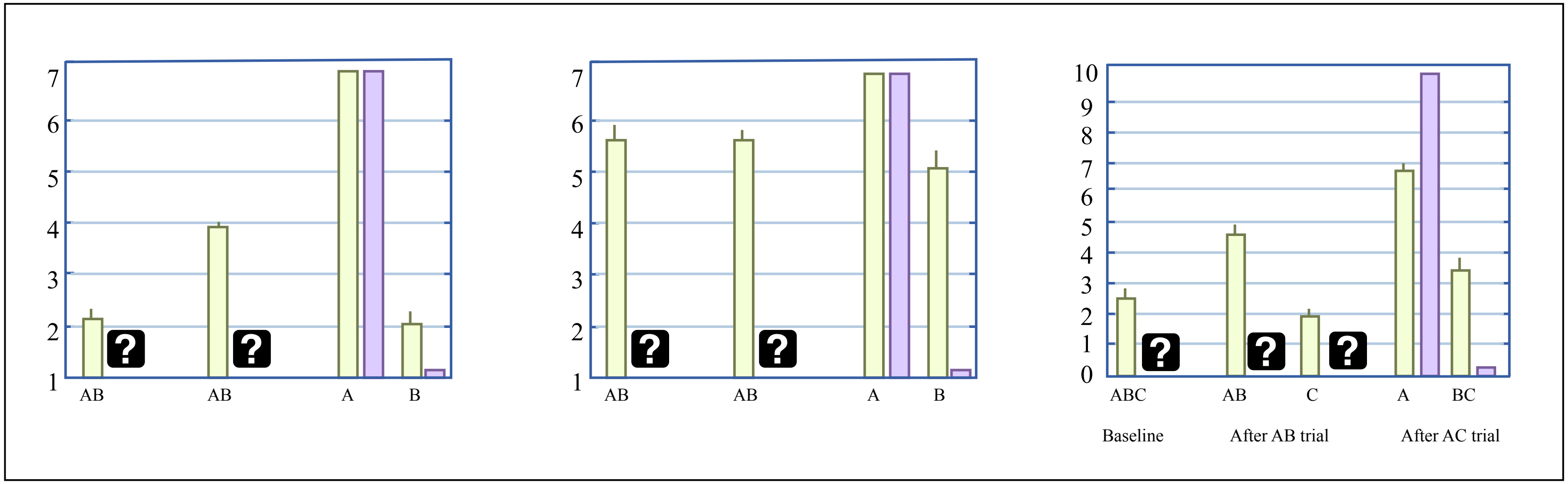


Figure by MIT OCW.

Bayes with correct theory:

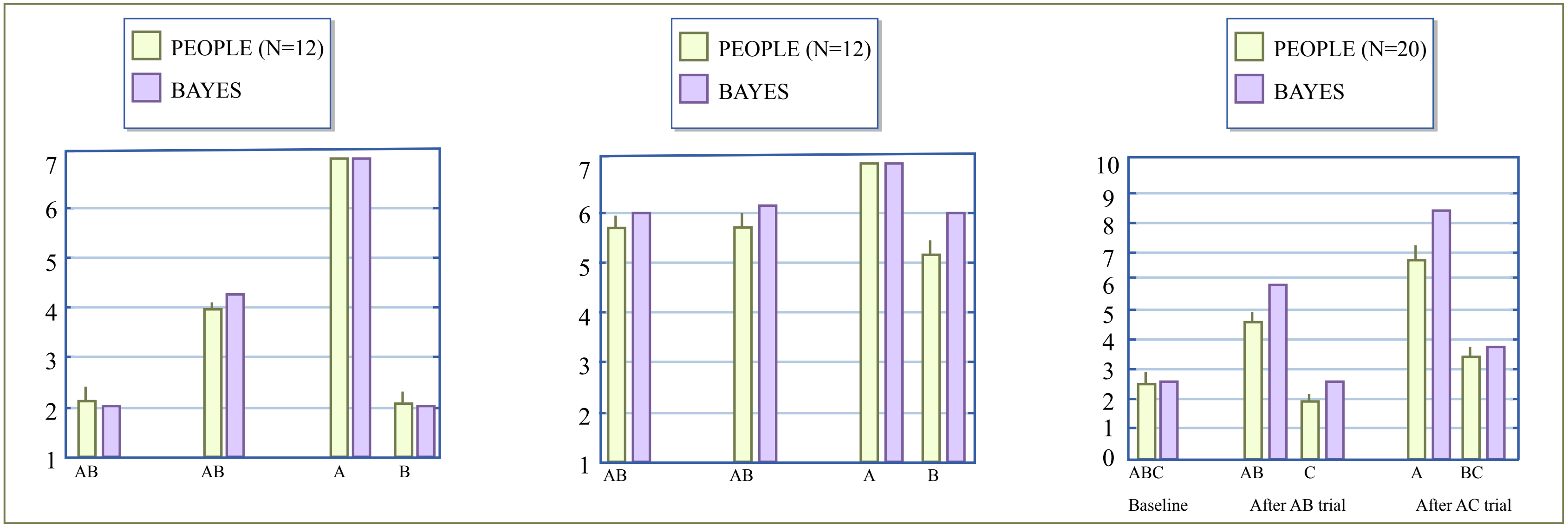


Figure by MIT OCW.

Bayes with "noisy sufficient causes" theory:

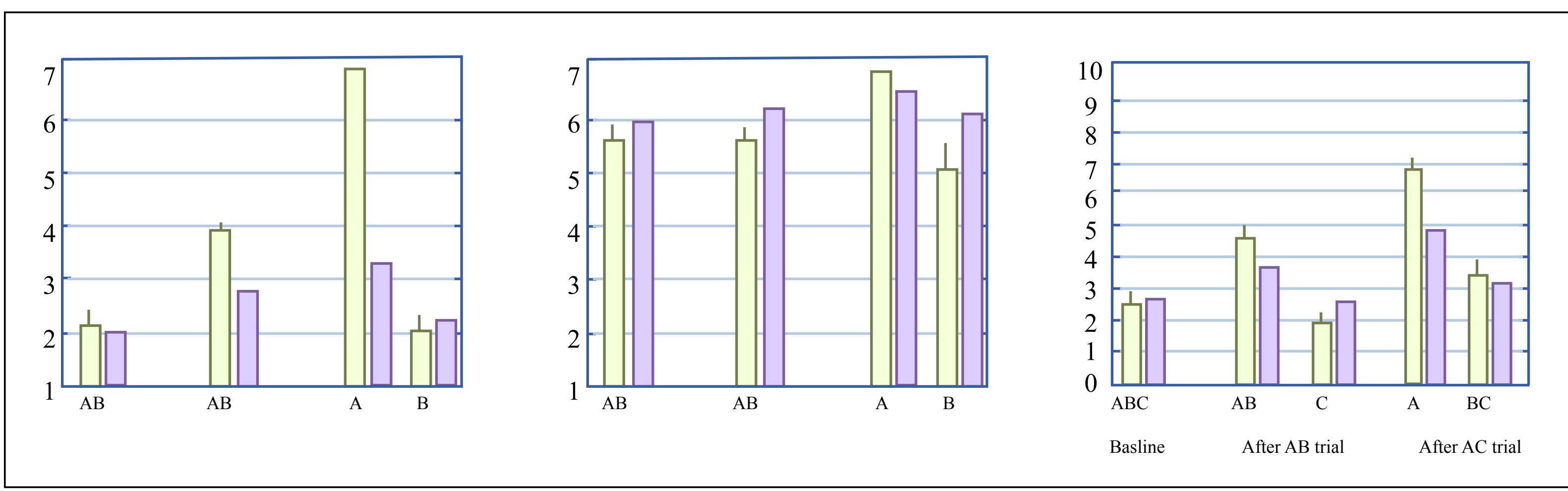


Figure by MIT OCW.

Blicket studies: summary

- Theory-based Bayesian approach explains one-shot causal inferences in physical systems.
- Captures a spectrum of inference:
  - Unambiguous data: adults and children make all-or-none inferences
  - Ambiguous data: adults and children make more graded inferences
- Extends to more complex cases with hidden variables, dynamic systems, ….