Model Selection and Parameter Estimation in Causal Learning

Hongjing Lu
Department of Psychology
University of Hong Kong

Bayesian Cognitive Science

- Bayes provides a *framework* for psychological theories

- Knowledge consists of *priors* and *generative models* (likelihoods)

- Psychological theories specify priors and generative models employed by humans

- Bayesian framework facilitates testing alternative psychological theories

Bayesian issues in causal learning

- What generating models do people use to interpret causal data?
- What priors guide causal learning?
- Can “logical” priors be formulated as probability distributions?

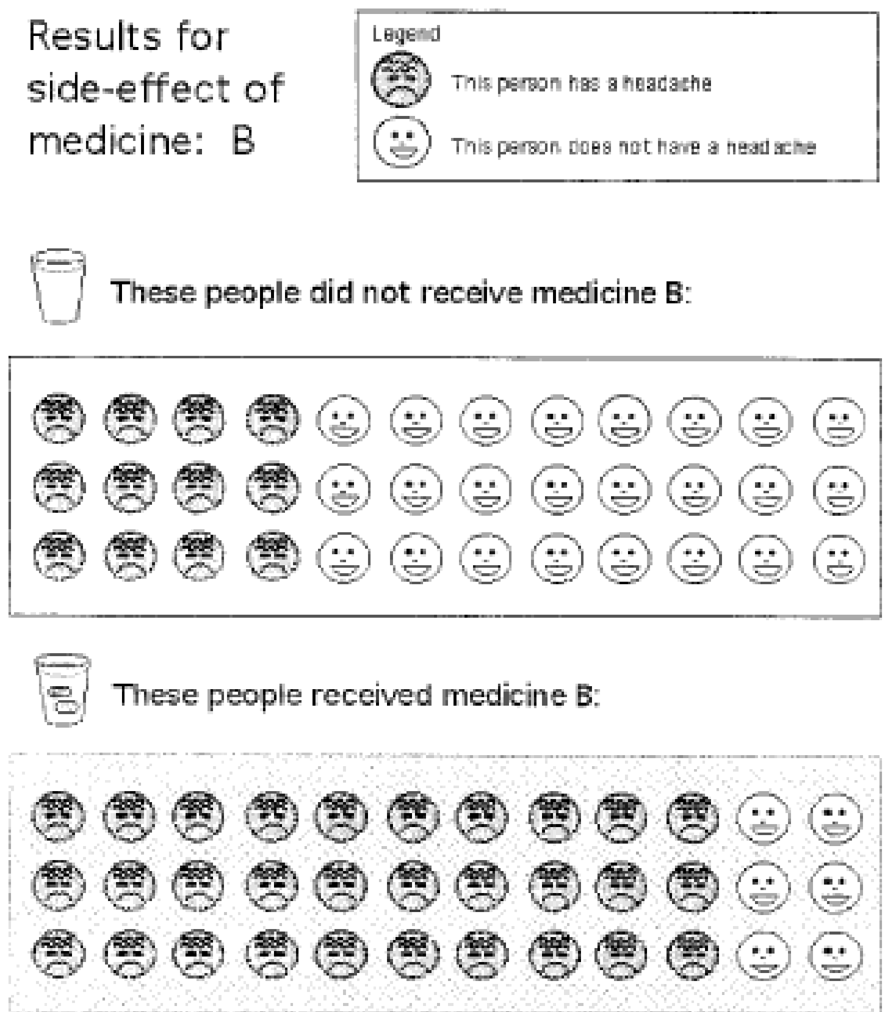
Outline

- Two main types of queries in human reasoning study
  - Structure judgment: model selection
  - Strength judgment: parameter estimation
- Generative model in causal reasoning
  - Linear model
  - Noisy logic model
- Generic prior in causal reasoning
  - Necessary & Sufficient prior

Elemental causal learning

12/36

30/36



Buehner, Cheng & Clifford (2004)

Causal structure with binary variables

Background cause, $B \in \{0,1\} = \{b^-, b^+\}$

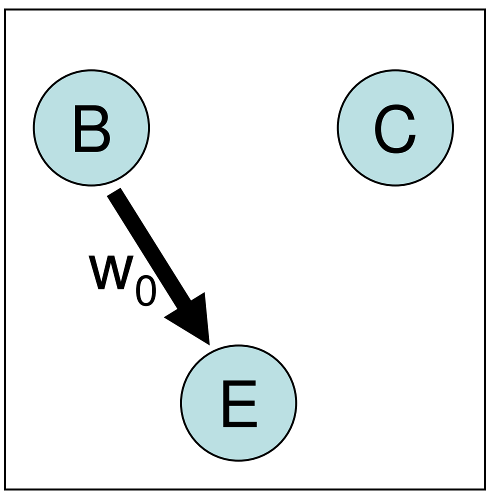
Candidate cause, $C \in \{0,1\} = \{c^-, c^+\}$

Effect, $E \in \{0,1\} = \{e^-, e^+\}$

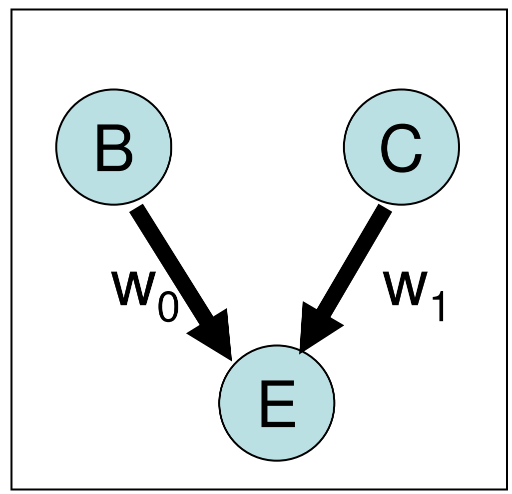
w_0 Probability of E occurring in the presence of **just** background cause

w_1 Probability of E occurring in the presence of **just** candidate cause

Graph 0



Graph 1



Griffiths & Tenenbaum, 2005

Causal learning experiment

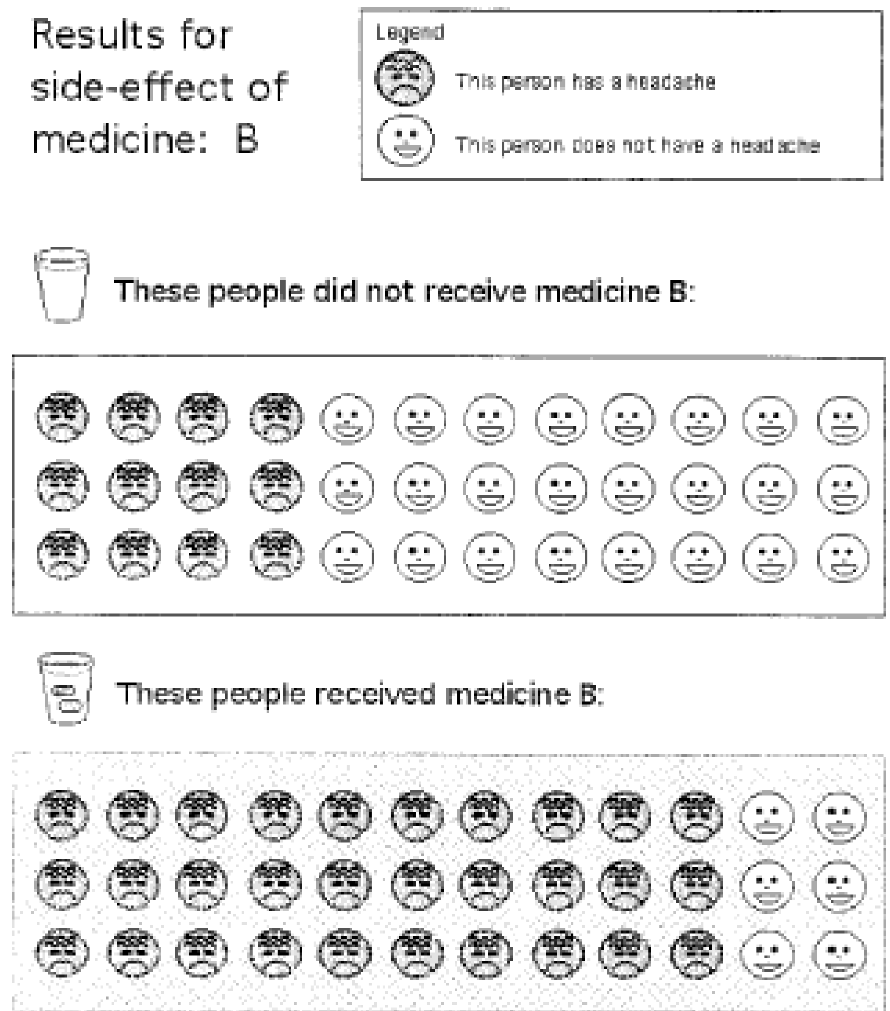
Causal direction:
generative

N(e+,c-) = 12

N(c-) = 36

N(e+,c-) = 30

N(c-) = 36



Two types of queries

- Contingency input
  - N(e+,c-), N(c-), N(e+,c+), N(c+)
- Structure judgment (Griffiths & Tenenbaum, 2005)
  - Does medicine B cause headaches?
  - How likely does medicine B cause headaches?

Two types of queries

- Contingency input
  - N(e+,c-), N(c-), N(e+,c+), N(c+)

- Strength judgment
  - Suppose that there are 100 people that **do not** have headaches. If this medicine was given to these 100 people, how many of them would have headaches?

Debate in causal learning

- Δ P (Jenkins & Ward, 1965; Rescorla, 1968)

$$\Delta P = P(e+ \mid c+) - P(e+ \mid c-)$$

- Causal power (Cheng, 1997)

$$q_G = \frac{\Delta P}{1 - P(e^+ \mid c^-)}$$ for generative cause

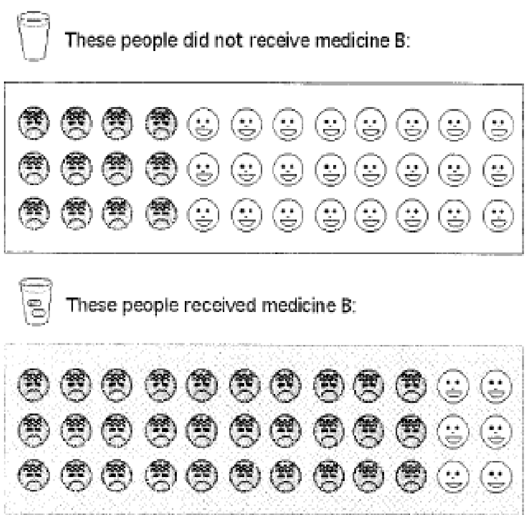
$$q_P = \frac{-\Delta P}{P(e^+ \mid c^-)}$$ for preventive cause

Bayesian view

- Point estimate using different generative models in causal learning (Griffiths & Tenenbaum, 2005)

- Maximum likelihood estimate

$$P(D \mid w_0, w_1) = P(N(e+,c-), N(c-) \mid w_0) P(N(e+,c+), N(c+) \mid w_0, w_1)$$



$$P(N(e+,c-), N(c-) \mid w_0) = \binom{N(c-)}{N(e+,c-)} P_1^{N(e+,c-)} (1-P_1)^{N(c-)-N(e+,c-)}$$

$$P_1 = P(e+ \mid b+, c-; w_0) = w_0$$

$$P(N(e+,c+), N(c+) \mid w_0, w_1) = \binom{N(c+)}{N(e+,c+)} P_2^{N(e+,c-)} (1-P_2)^{N(c-)-N(e+,c-)}$$

$$P_2 = P(e+ \mid b+, c+; w_0, w_1) = w_0$$

Generative model in causal learning

- Linear model

$$P(e^+ \mid b, c; w_0, w_1) = w_0 b + w_1 c \quad \text{with the constraint } \omega_0 + \omega_1 \le 1$$

- Noisy logic model
  - Generative cause: noisy-or model, adding probability on the logical OR function that the effect occurs if and only if B or C are present

$$P(e^+ \mid b, c; w_0, w_1) = 1 - (1 - w_0 b)(1 - w_1 c)$$

  - Preventive cause: noisy-and-not model, adding probability on the logical AND-NOT function that E will occur if B occurs and not C

$$P(e^+ \mid b, c; w_0, w_1) = w_0 b(1 - w_1 c)$$

MLE with linear model

$$P(D \mid w_0, w_1) = \binom{n}{a} {w_0}^{a} (1 - w_0)^{n-a} \binom{n}{b} (w_0 + w_1)^{b} (1 - w_0 - w_1)^{n-b}$$

$$where\ \ a = N(e+, c-),\ \ b = N(e+, c+),\ \ n = N(c-) = N(c+)$$

$$F(w_0, w_1) = \log P(D \mid w_0, w_1)$$

$$\frac{\partial F}{\partial w_0} = 0 \qquad \frac{a}{w_0} - \frac{n-a}{1-w_0} + \frac{b}{w_0 + w_1} - \frac{n-b}{1 - w_0 - w_1} = 0 \qquad w_0 = \frac{a}{n} = P(e+ \mid c-)$$

$$\Rightarrow \qquad \Rightarrow$$

$$\frac{\partial F}{\partial w_1} = 0 \qquad \frac{b}{w_0 + w_1} - \frac{n-b}{1 - w_0 - w_1} = 0 \qquad w_1 = \frac{b-a}{n} = P(e+ \mid c+) - P(e+ \mid c-)$$

MLE with noisy-or model

$$P(D \mid w_0, w_1) = \binom{n}{a} w_0{}^a (1-w_0)^{n-a} \binom{n}{b} (1-(1-w_0)(1-w_1))^b ((1-w_0)(1-w_1))^{n-b}$$

$$F(w_0, w_1) = \log P(D \mid w_0, w_1)$$

$$\frac{\partial F}{\partial w_0} = 0 \qquad \frac{a}{w_0} - \frac{n-a}{1-w_0} + \frac{b}{w_0 + w_1 - w_0 w_1} - \frac{n-b}{1-w_0} = 0$$

$$\Rightarrow \qquad \Rightarrow$$

$$\frac{\partial F}{\partial w_1} = 0 \qquad \frac{b(1-w_0)}{w_0 + w_1 - w_0 w_1} - \frac{n-b}{1-w_1} = 0$$

$$w_0 = \frac{a}{N} = P(e+ \mid c-)$$

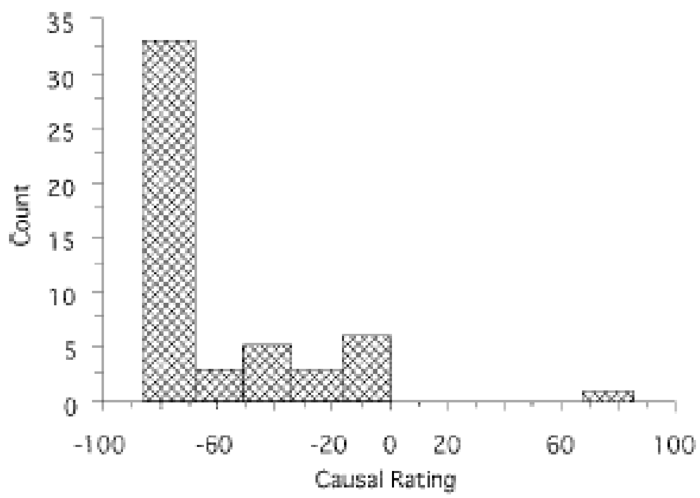
$$w_1 = \frac{b-a}{n-a}$$

$$= \frac{P(e+ \mid c+) - P(e+ \mid c-)}{1 - P(e+ \mid c-)}$$

Uncertainty in strength judgment

Table 3
Design and Results of Experiment 2

Condition	Power	ΔP	P(e|c)	P(e|~c)	Causal ratings M (SD)	Causal ratings Mdn
A	1.00	0.50	36/36	18/36	85.7 (26.5)	100
B	0.75	0.50	30/36	12/36	67.8 (19.1)	75
C	0.75	0.75	27/36	0/36	74.4 (5.7)	75
D	0.50	0.50	18/36	0/36	48.5 (14.1)	50
E	0.75	−0.50	6/36	24/36	−59.7 (31.4)	−75
F	0.75	−0.75	9/36	36/36	−66.5 (18.8)	−75
G	0.50	−0.50	18/36	36/36	−44.5 (21.1)	−50
H	0.00	0.00	12/36	12/36	−0.5 (4.86)	0
I	0.00	0.00	24/36	24/36	0.7 (3.78)	0
J	1.00	−0.50	0/36	18/36	−85.9 (26.6)	−100



Buehner, Cheng & Clifford (2004)

Strength MAP estimate

- MAP estimate

$$P(w_1 \mid D) = \int_0^1 P(w_0, w_1 \mid D)\, dw_0 = \int_0^1 \frac{P(D \mid w_0, w_1) P(w_0, w_1)}{P(D)}\, dw_0$$

- Generative model $P(D \mid w_0, w_1)$
  - Linear model
  - Noisy logic model
- Prior $P(w_0, w_1)$
  - Uniform
  - Generic prior in causal learning ??

Generic prior
Necessity & Sufficiency

- A possible generic prior for reasoning: necessary and sufficient cause (J. S. Mill, 1843)

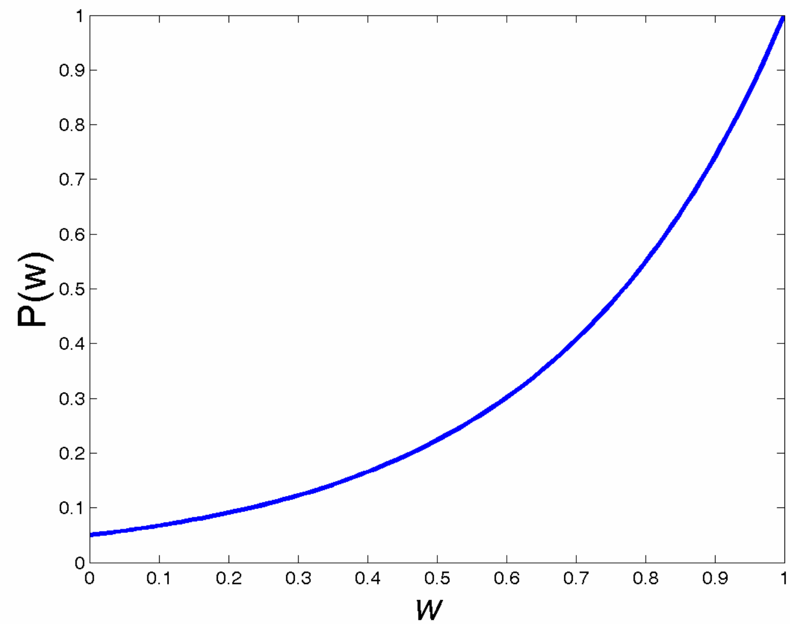
- Example: A genetic defect on chromosome 4 is necessary and sufficient to cause Huntington's disease
  - **<u>Sufficiency</u>**: Huntington's disease occurs as long as an individual has the genetic defect
  - **<u>Necessity</u>**: Huntington's disease never occurs in the absence of the genetic defect

- Generic prior is **probabilistic**, but biased toward a necessary and sufficient cause

Generic prior
Necessity & Sufficiency

- Preference for sufficient cause
  - Higher causal strength better

$$P(w) \propto e^{-\alpha(1-w)}$$



Generic prior
Necessity & Sufficiency

- Preference for necessity
  - Causal simplicity (Chater & Vitányi, 2003) potentially could manifest itself in multiple ways, which likely include a preference for fewer causes and for causes that do not involve interactions

Prior for simplicity

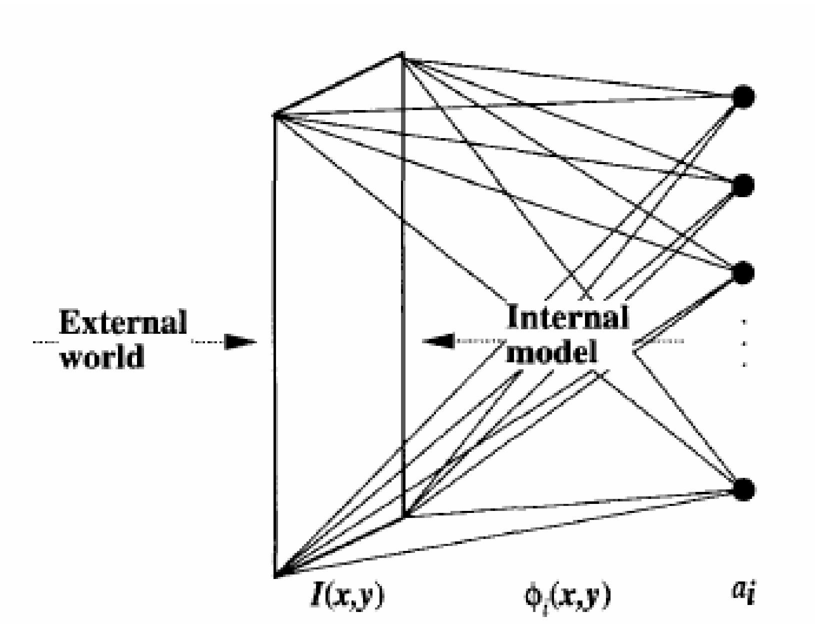
- Sparse coding (Olshausen & Field, 1996)

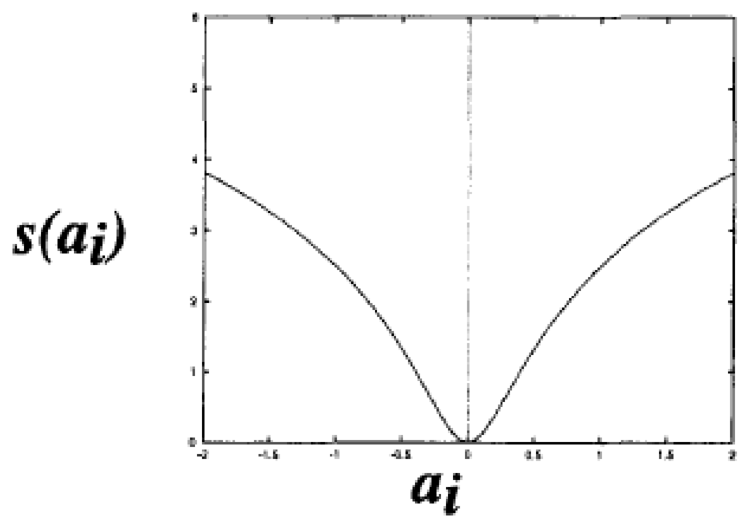
$$I(x, y) = \sum_i a_i \phi_i(x, y)$$

$$P(I \mid \phi) = \int P(I \mid a, \phi) P(a) da$$

Sparseness prior

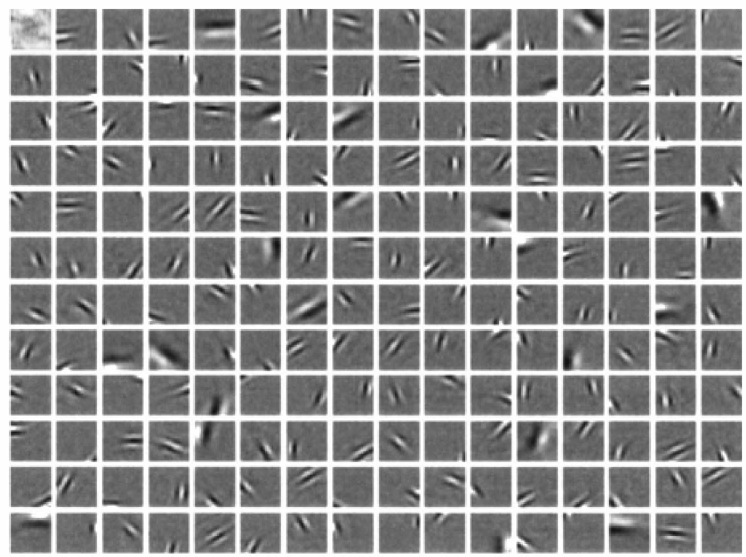
$$P(a) \propto e^{-\beta \sum_i S(a_i)} = e^{-\beta \sum_i \left| \frac{a_i}{\sigma} \right|}$$





Prior for simplicity

- Learned basis functions using sparseness prior

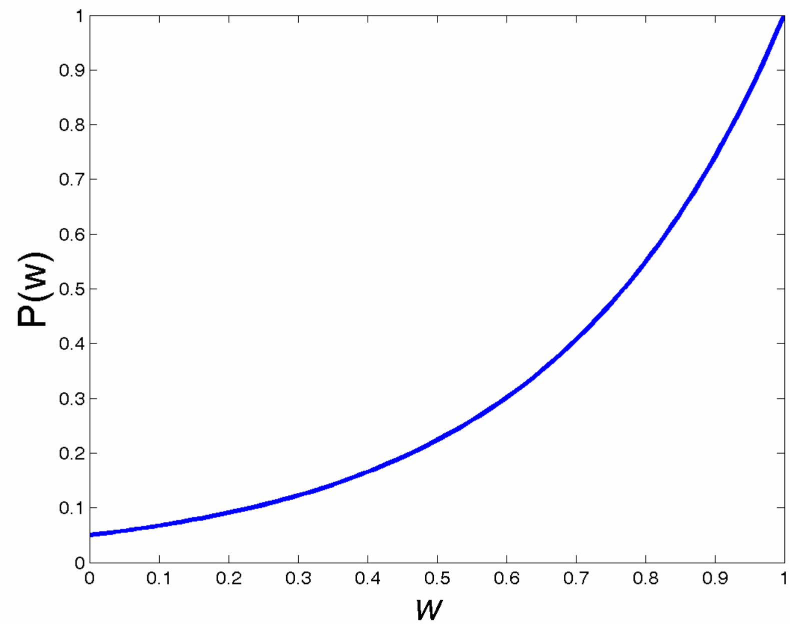


Sparseness prior $P(a) \propto e^{-\beta \sum_i S(a_i)} = e^{-\beta \sum_i \left| \frac{a_i}{\sigma} \right|}$

Generic prior
Necessity & Sufficiency

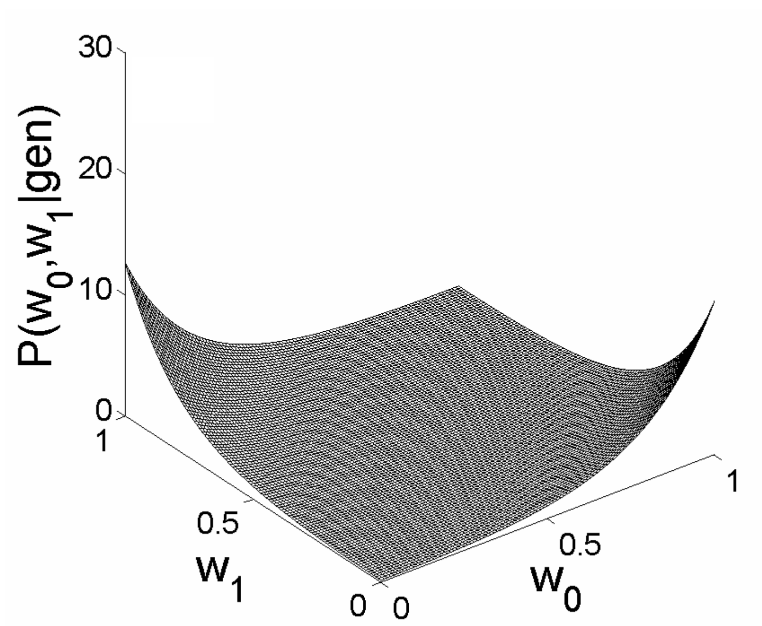
- Preference for sufficient cause
  - Higher causal strength better

$$P(w) \propto e^{-\alpha(1-w)}$$



Generic prior
Necessity & Sufficiency

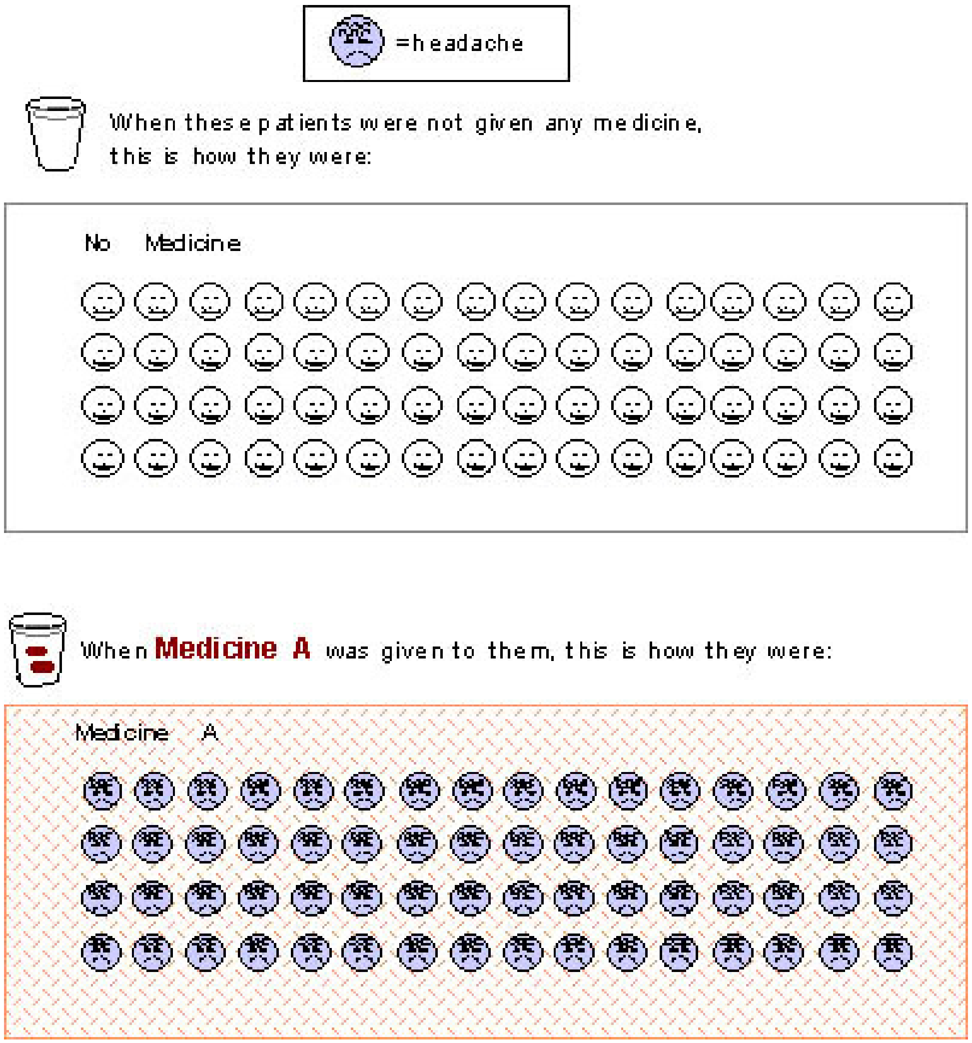
- Use mixed distribution to model the preference for necessary & sufficient cause



$$P(w_0, w_1 \mid Graph1)$$
$$\propto e^{-\alpha w_0 - \alpha(1-w_1)} + e^{-\alpha(1-w_0) - \alpha w_1}$$

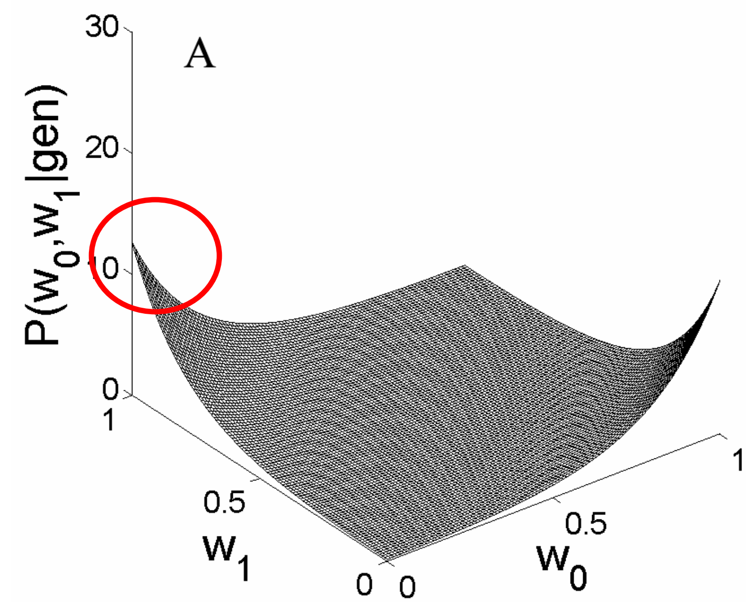
Lu, Yuille, Liljeholm, Cheng & Holyoak, *Cognitive Science Proceedings*, 2006, 2007

A favorable condition for generative cause



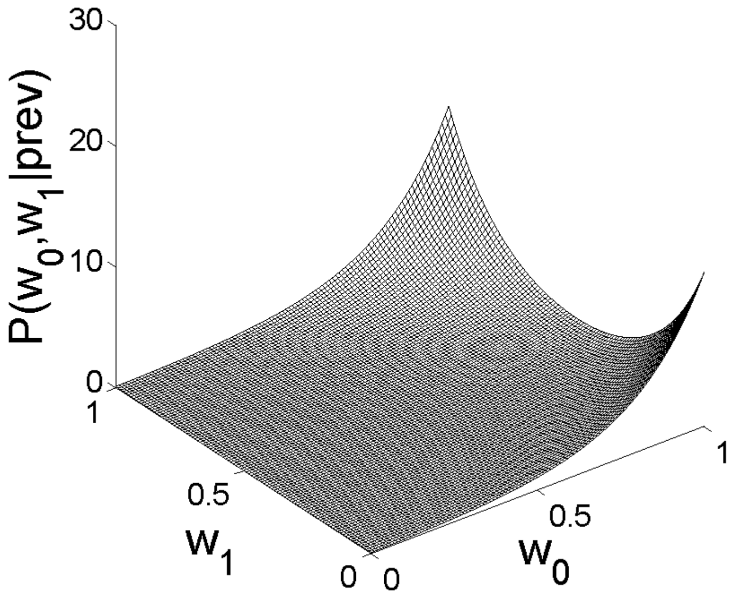
Medicine A produces headache

$W_0=0, W_1=1$



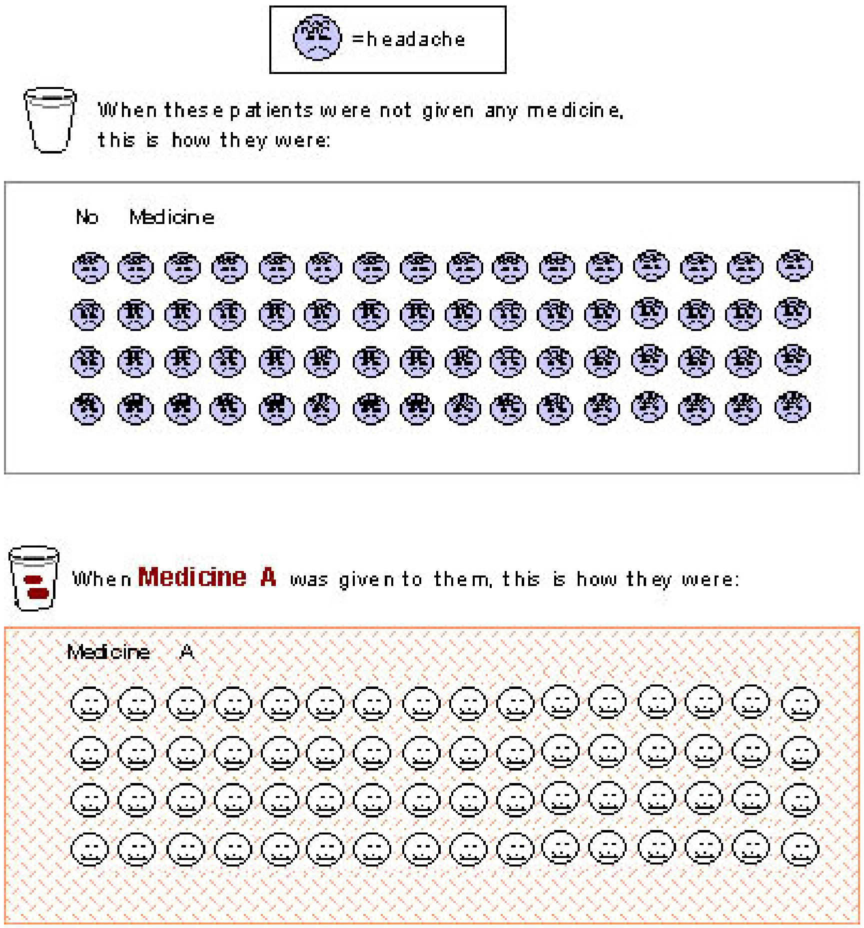
Generic prior
Necessity & Sufficiency

- Preventive case



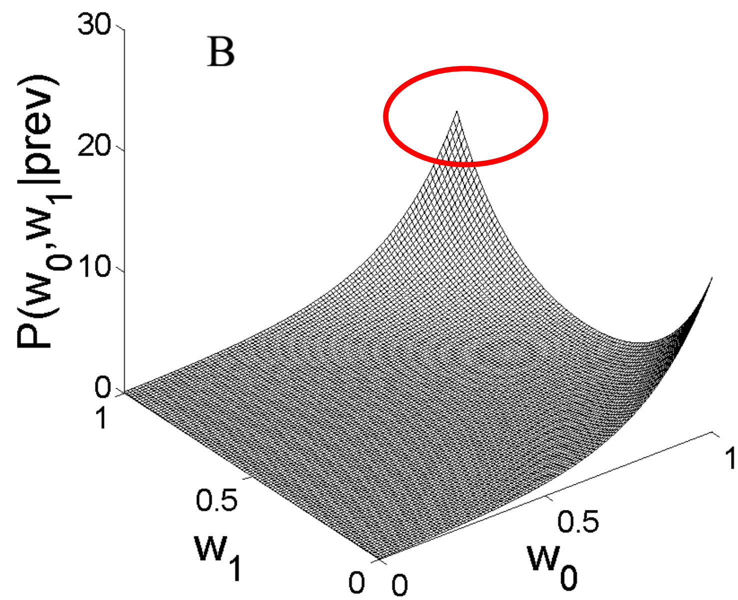
$$P(w_0, w_1 \mid Graph1) \propto e^{-\alpha(1-w_0)-\alpha(1-w_1)} + e^{-\alpha(1-w_0)-\alpha w_1}$$

Favorable Conditions in Preventive Cause



Medicine A prevents headache

$W_0=1,\ W_1=1$



Strength MAP estimate

- MAP estimate

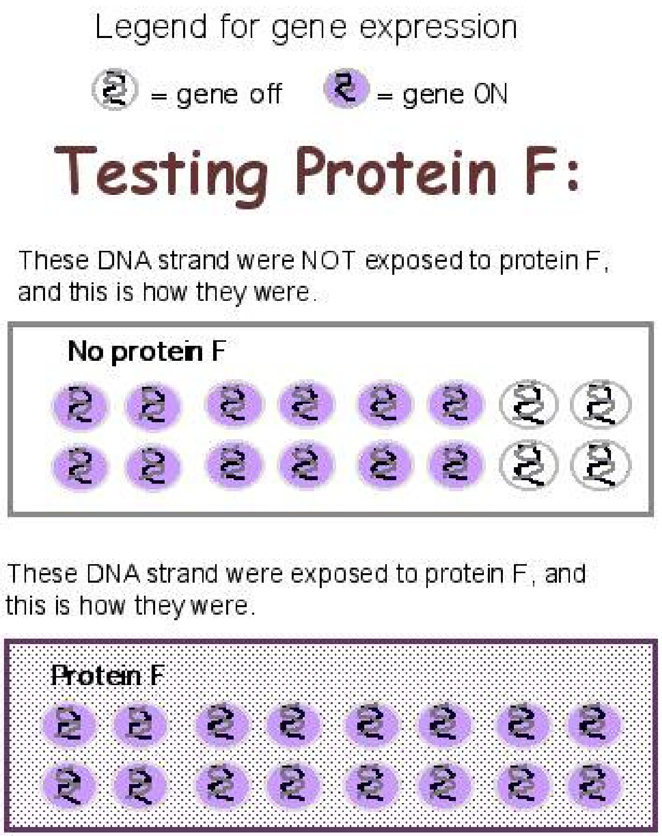
$$P(w_1 \mid D) = \int_0^1 P(w_0, w_1 \mid D)\,dw_0 = \int_0^1 \frac{P(D \mid w_0, w_1) P(w_0, w_1)}{P(D)}\,dw_0$$

- Generative model $P(D \mid w_0, w_1)$
  - Noisy logic model
- Prior $P(w_0, w_1)$
  - Uniform
  - NS prior

Causal strength experiment
(Liljeholm & Cheng, 2006)

- Evaluate the influence of protein on the expression of a gene
- Each participant ran 8 different contingency conditions with the same causal direction
- Participants were informed of causal direction, generative or preventive cause

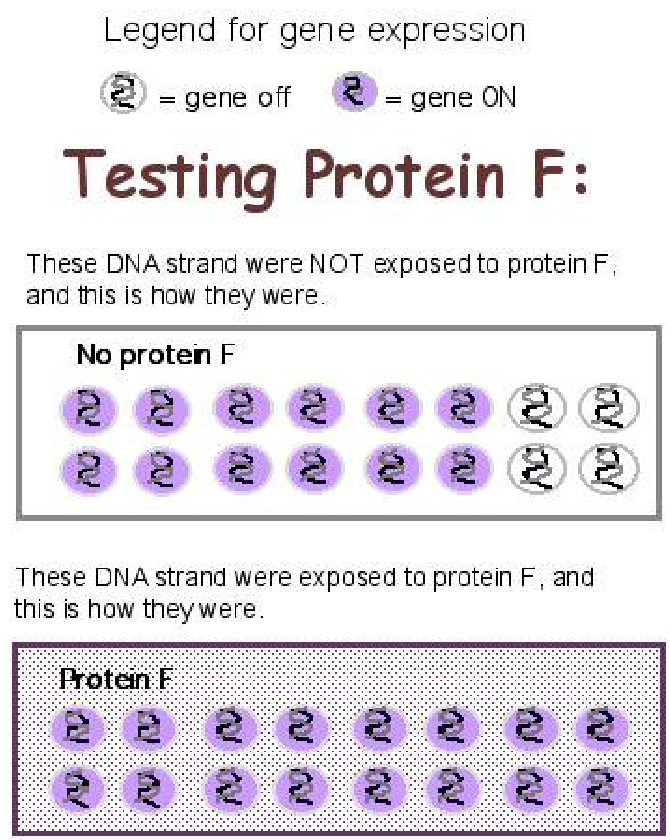
Causal strength experiment
(Liljeholm & Cheng, 2006)



Suppose that there is a sample of 100 DNA strands and that the gene is OFF in all those DNA strands. If these 100 strands were exposed to the protein, in how many of them would the gene be TURNED ON?

Causal strength experiment
(Liljeholm & Cheng, 2006)

Causal direction: generative



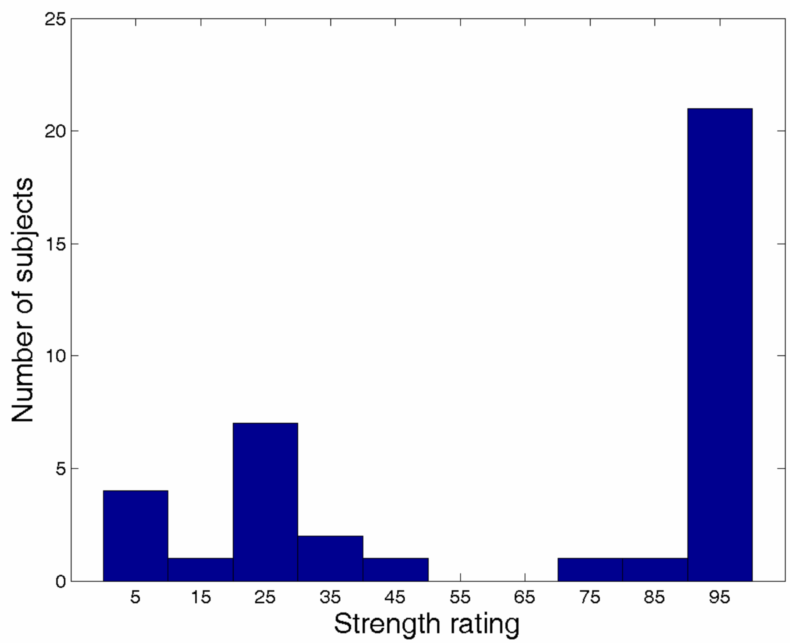
P(e+|c-) = 12/16

P(e+|c+) = 16/16

Causal strength experiment
(Liljeholm & Cheng, 2006)

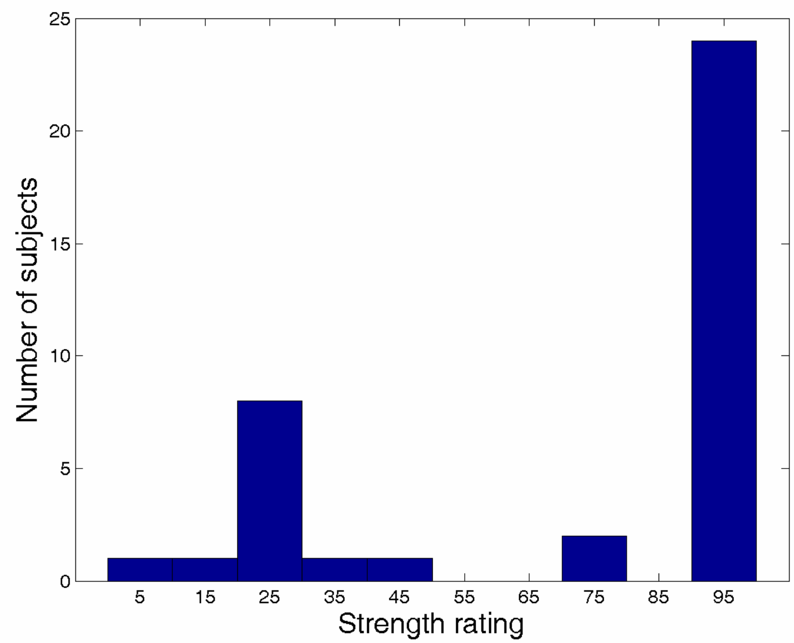
P(e+|c-) = 12/16,   P(e+|c+)=16/16

Power = 1;   ΔP=0.25



P(e+|c-) = 48/64,   P(e+|c+)=64/64

Power = 1;   ΔP=0.25



Causal strength experiment

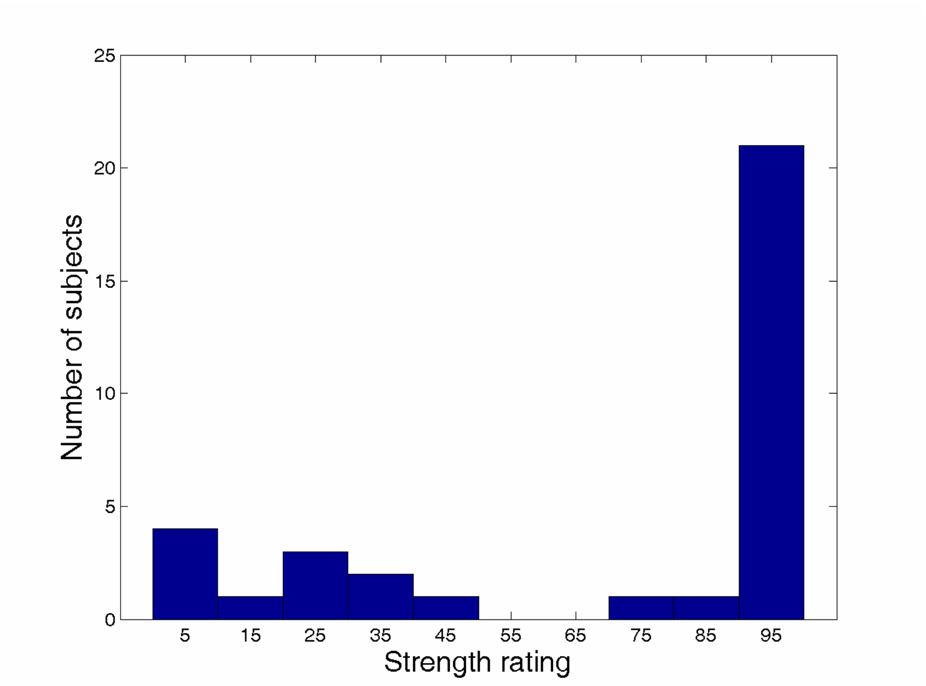
- We find that, across all 8 contingency conditions, four subjects out of thirty-eight consistently used P(e+|c+) -P(e+|c-) to rate the strength

- We excluded these four subjects in histogram fitting

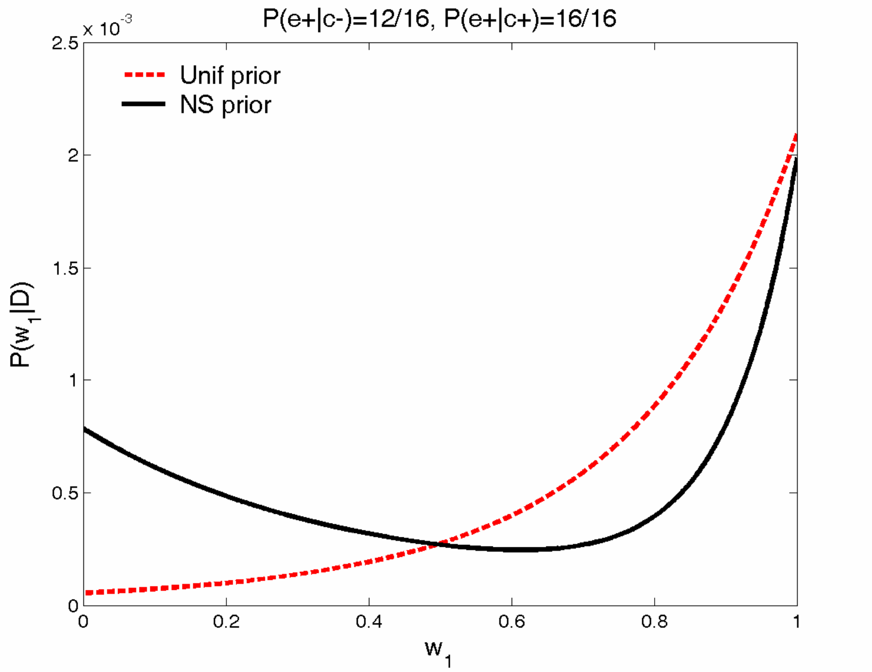
Causal strength experiment

P(e+|c-) = 12/16,   P(e+|c+)=16/16, Power = 1;   ΔP=0.25

Human



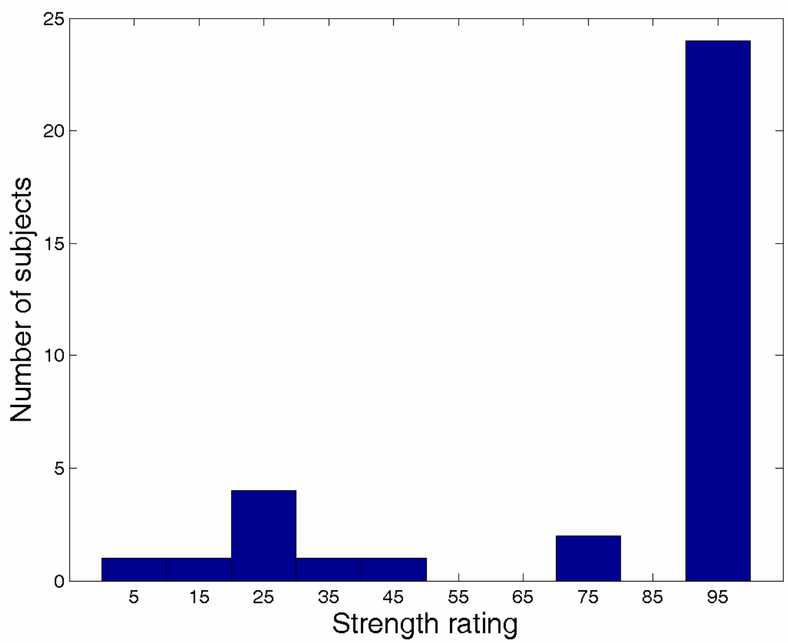
Model



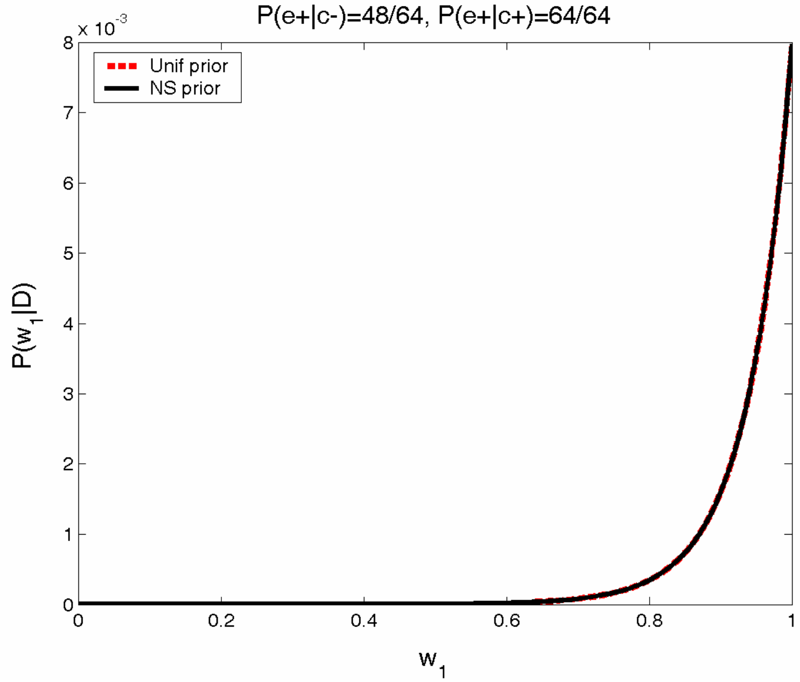
Causal strength experiment

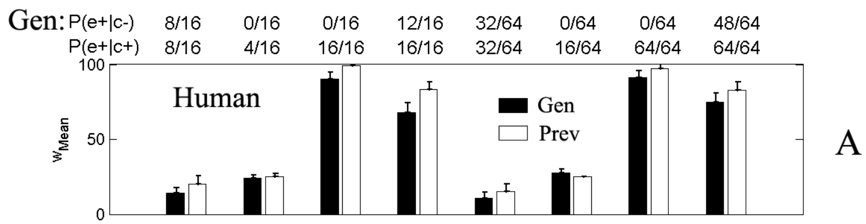
P(e+|c-) = 48/64,   P(e+|c+)=64/64, Power = 1; Δ P=0.25

Human



Model





Prev: P(e+|c-) 8/16 16/16 16/16 4/16 32/64 64/64 64/64 16/64
P(e+|c+) 8/16 12/16 0/16 0/16 32/64 48/64 0/64 0/64

Two types of queries

- Contingency input
  - N(e+,c-), N(c-), N(e+,c+), N(c+)
- Structure judgment
  - How likely does medicine B cause headaches?
  - Does medicine B cause headaches?

Causal structure

- Causal support with generic prior

$$\frac{P(Graph\,1 \mid D)}{P(Graph\,0 \mid D)} = \frac{P(D \mid Graph\,1)P(Graph\,1)}{P(D \mid Graph\,0)P(Graph\,0)}$$

$$= \frac{\int_0^1 \int_0^1 P(D \mid w_0, w_1, Graph\,1) P(w_0, w_1 \mid Graph\,1)\, dw_0\, dw_1}{\int_0^1 P(D \mid w_0, Graph\,0) P(w_0 \mid Graph\,0)\, dw_0} \frac{P(Graph\,1)}{P(Graph\,0)}$$

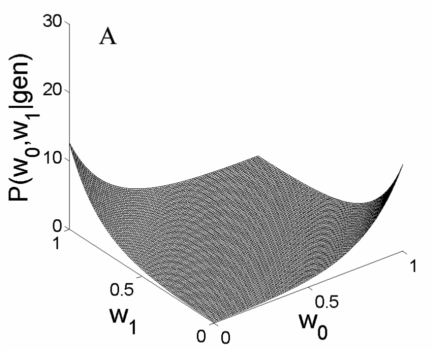
Likelihood

Prior

Generic prior
Necessity & Sufficiency

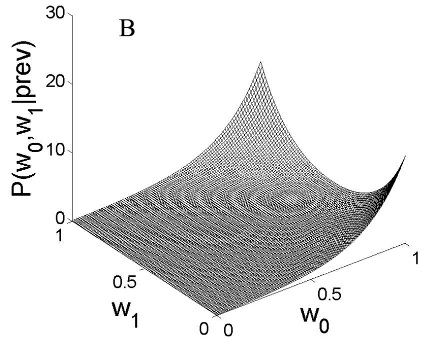
- Use mixed distribution to model the preference for necessary & sufficient cause

Generative case



$$P(w_0, w_1 \mid Graph1) \propto e^{-\alpha w_0 - \alpha(1-w_1)} + e^{-\alpha(1-w_0)-\alpha w_1}$$

Preventive case

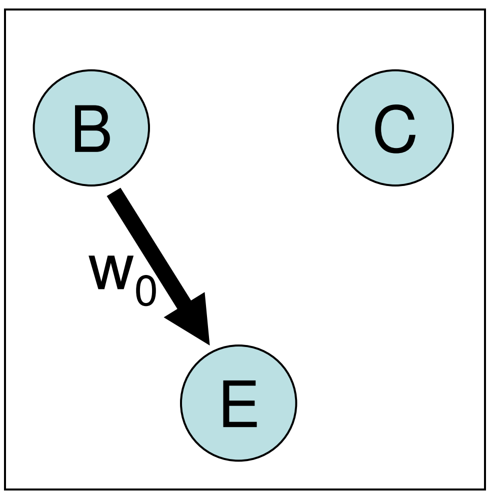


$$P(w_0, w_1 \mid Graph1) \propto e^{-\alpha(1-w_0) - \alpha(1-w_1)} + e^{-\alpha(1-w_0)-\alpha w_1}$$

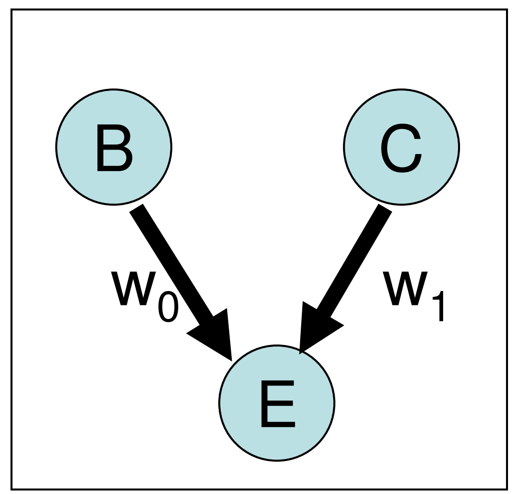
Human model selection favors sufficiency of candidate cause

- New candidate cause accepted only if (relatively) sufficient (i.e., high strength)
- Consistent with limitations on working memory
- if only a limited number of causes can be considered, best to focus on strong causes

Graph 0

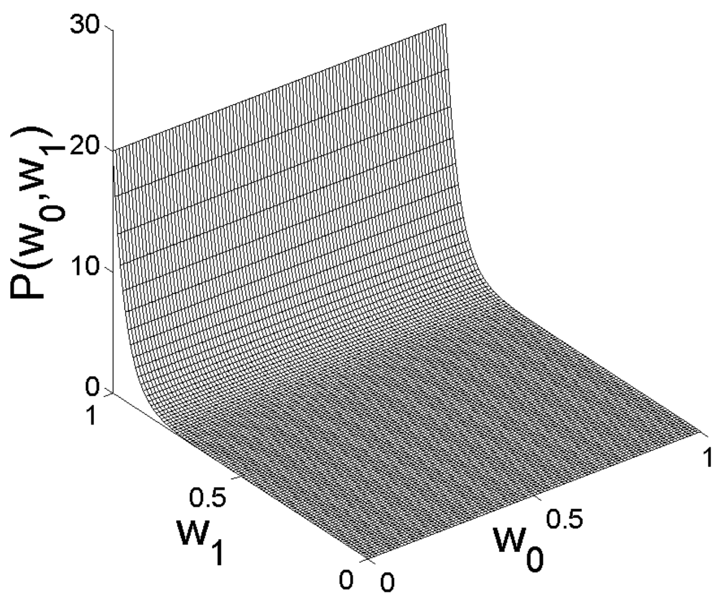


Graph 1



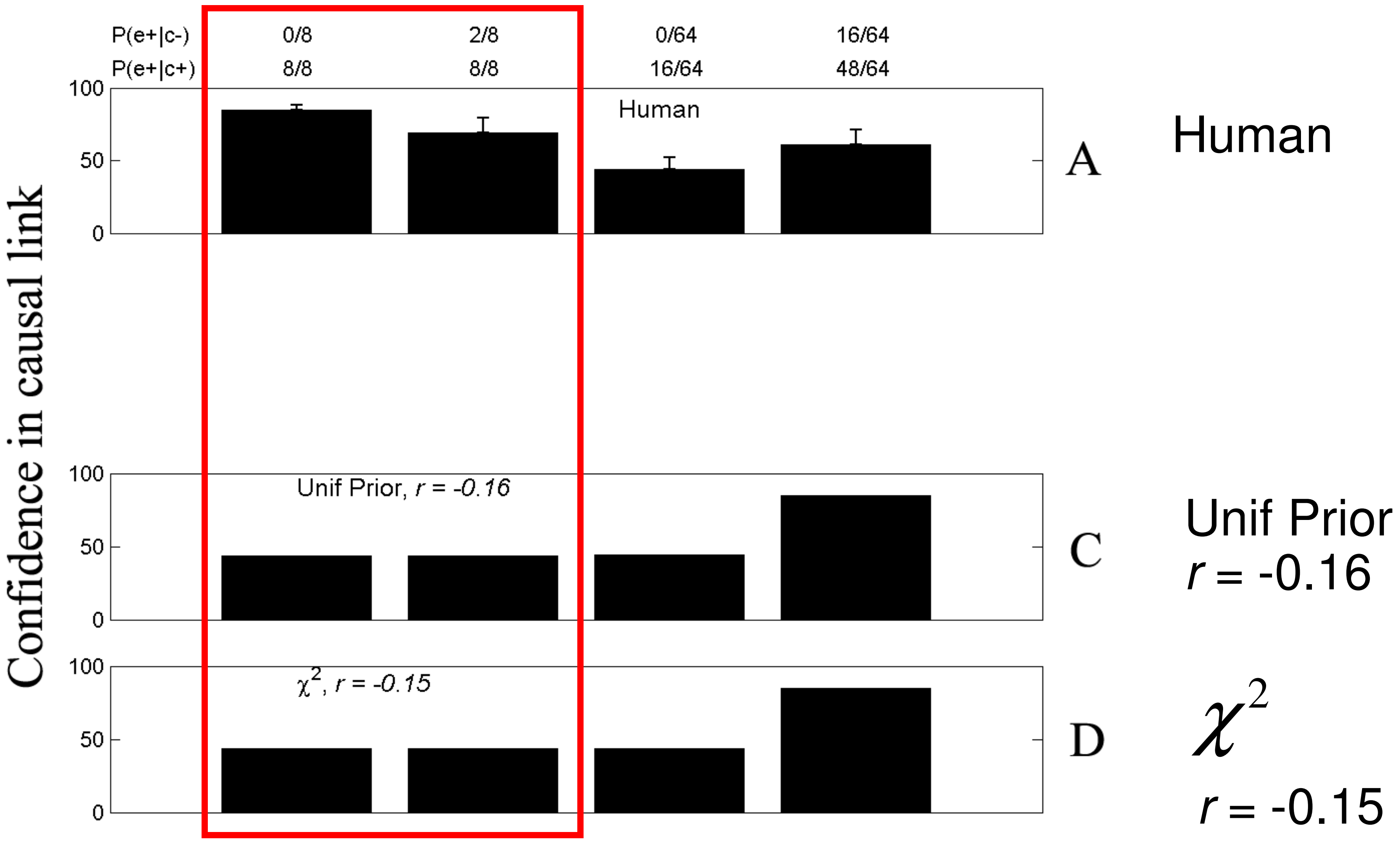
Generic prior for structure judgment
Sufficiency of C

- C accepted as a new cause only if it is a strong cause

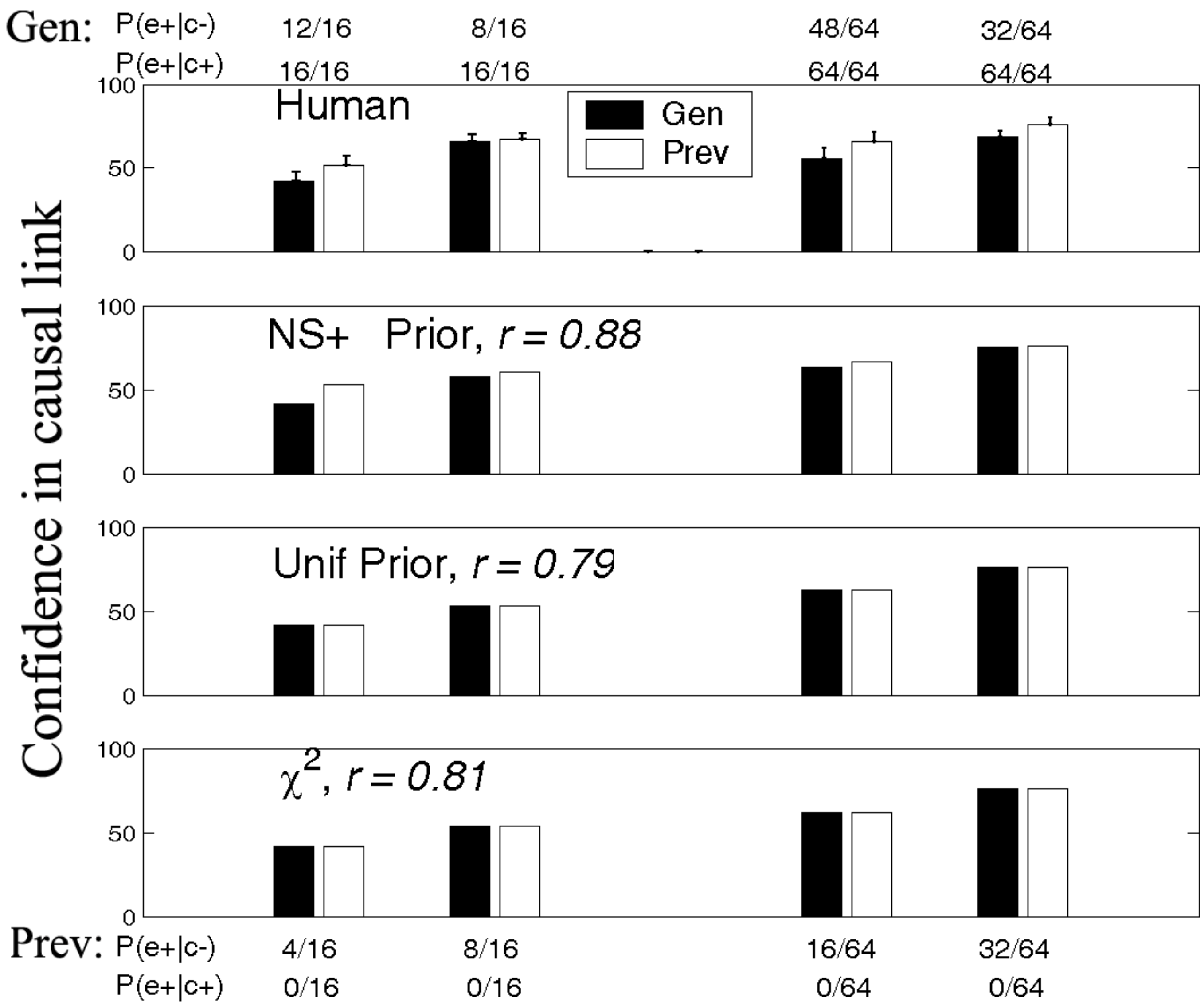


$$P(w_0, w_1) \propto e^{-\beta(1-w_1)}$$

Priors matter most when data is sparse



Generative and preventive asymmetry



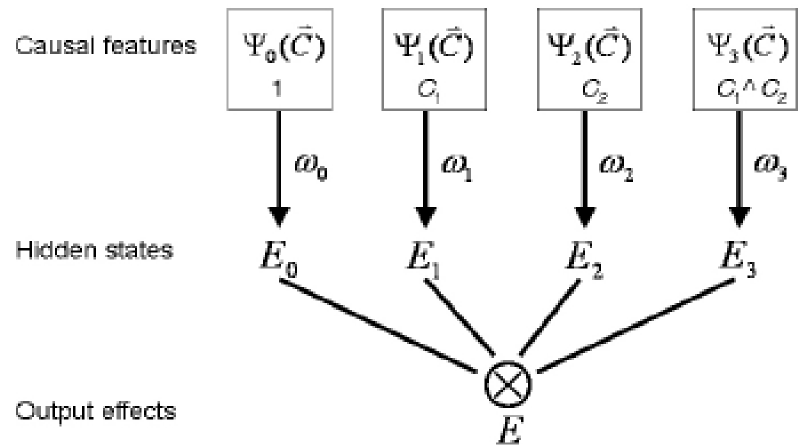
Human

NS Prior
r = 0.88

Unif Prior
r = 0.79

χ^2
r = 0.81

Noisy logic generative model



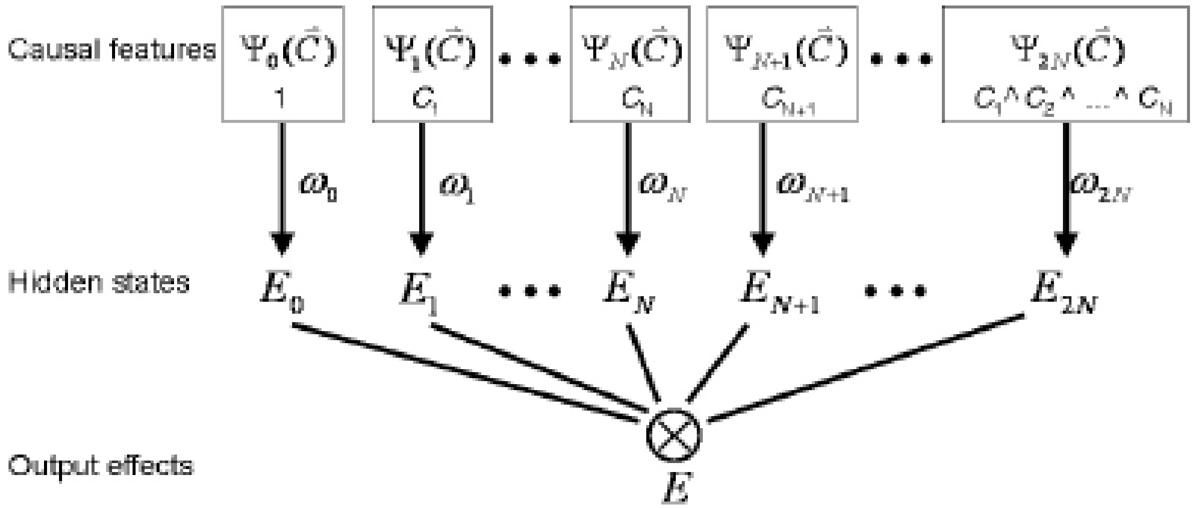
Noisy-or

$$
\begin{aligned}
P_{nl}(E=1|C_1,C_2;\omega_1,\omega_2) &= \sum_{E_1,E_2} \delta_{1,E_1 \vee E_2} P(E_1|\Psi_1(\vec{C});\omega_1)P(E_2|\Psi_2(\vec{C});\omega_2) \\
&= \omega_1 C_1(1-\omega_2 C_2) + (1-\omega_1 C_1)\omega_2 C_2 + \omega_1\omega_2 C_1 C_2 \\
&= \omega_1 C_1 + \omega_2 C_2 - \omega_1\omega_2 C_1 C_2 = P_{nor}(E=1|C_1,C_2;\omega_1,\omega_2) \quad (1)
\end{aligned}
$$

Noisy-and-not

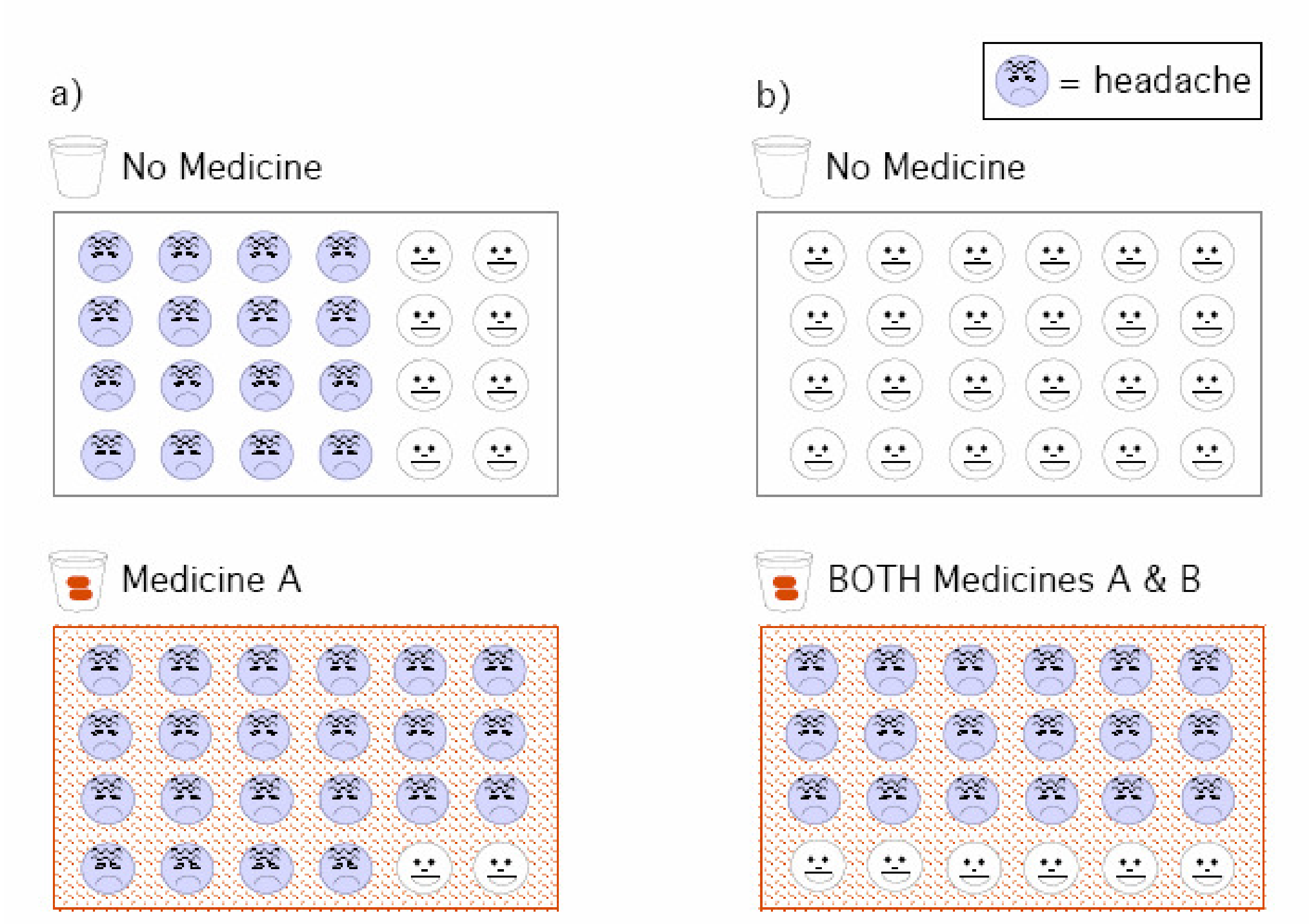
$$
\begin{aligned}
P_{nl}(E=1|C_1,C_2;\omega_1,\omega_2) &= \sum_{E_1,E_2} \delta_{1,E_1 \wedge \neg E_2} P(E_1|\Psi_1(\vec{C});\omega_1)P(E_2|\Psi_2(\vec{C});\omega_2) \\
&= \omega_1 C_1\{1-\omega_2 C_2\} = P_{n-and-not}(E=1|C_1,C_2;\omega_1,\omega_2). \quad (2)
\end{aligned}
$$

Generalized form



$$P(E=1|\vec{C};\vec{\omega}) = \sum_{\vec{E}} \delta_{E,f(E_0,\ldots,E_{2^N-1})} \prod_{i=0}^{2^N-1} P(E_i=1|\Psi_i;\omega_i).$$

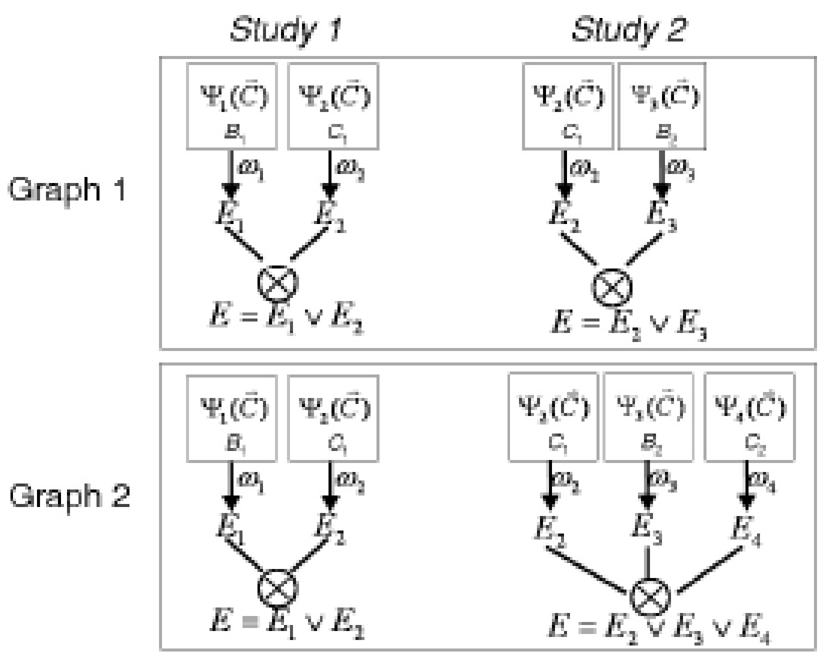
Causal generalization experiment
(Liljeholm & Cheng, *Psych Science*, in press)

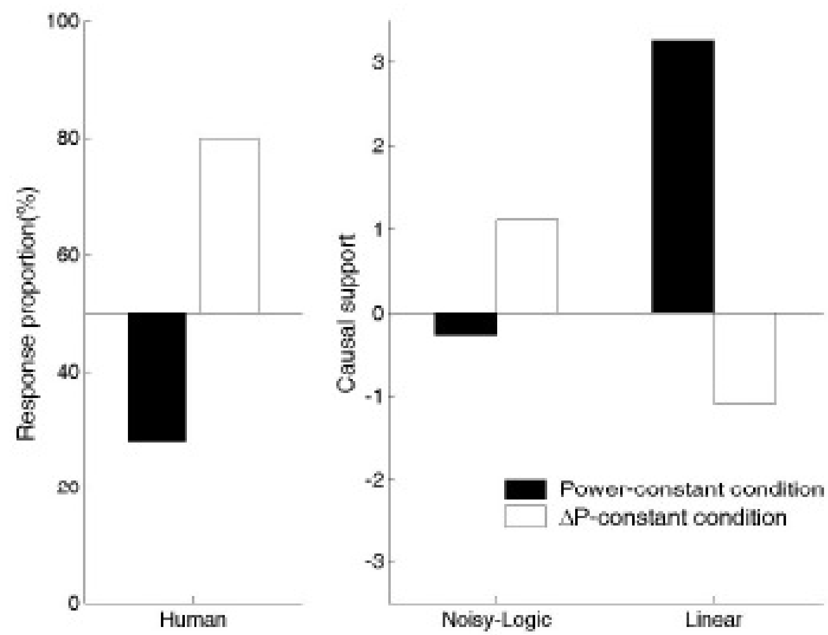


Based on the information from BOTH experiments, what is your best bet on whether or not Medicine B causes headache?

Causal generalization experiment

- Model selection





Causal interaction

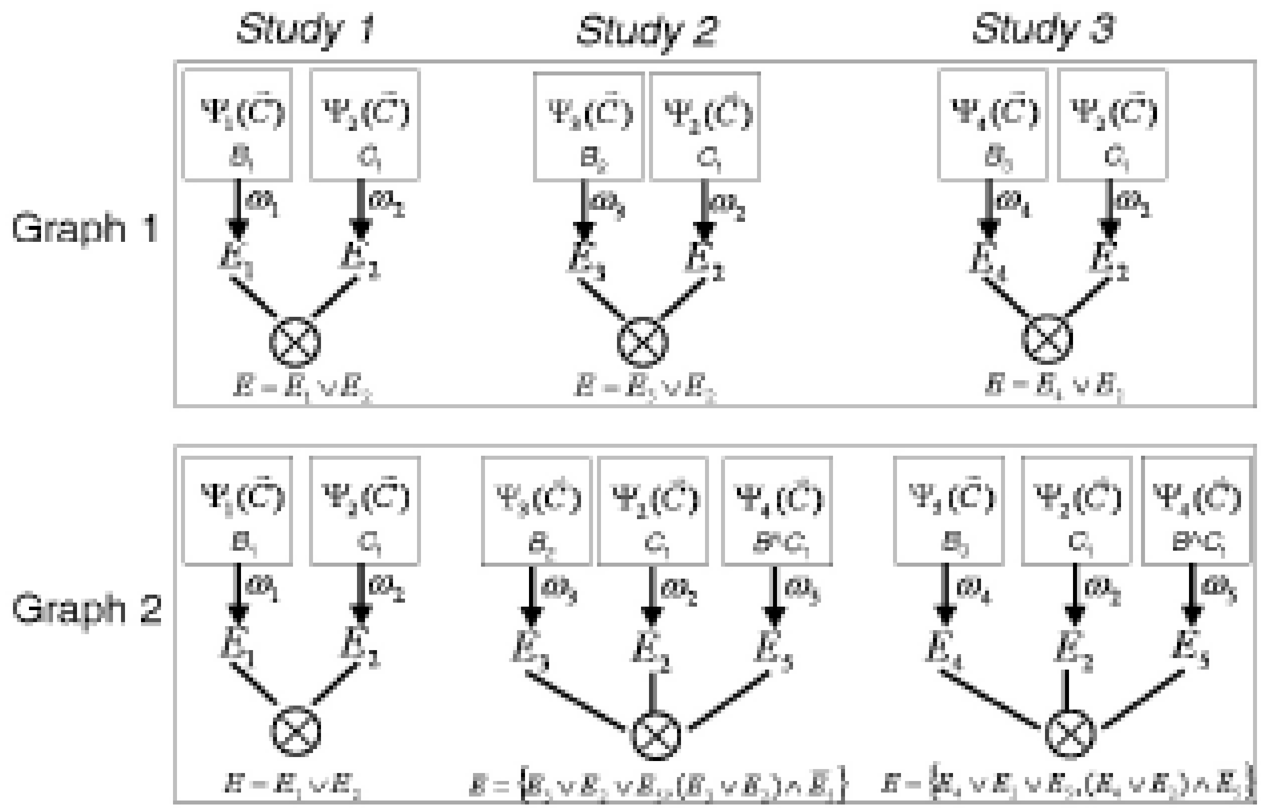
(Liljeholm & Cheng, in press)

	Study 1		Study 2		Study 3	
	e|no A	e|A	e|no A	e|A	e|no A	e|A
Power-Constant	16/24	22/24	8/24	20/24	0/24	18/24
Power-Varying	0/24	6/24	0/24	12/24	0/24	18/24

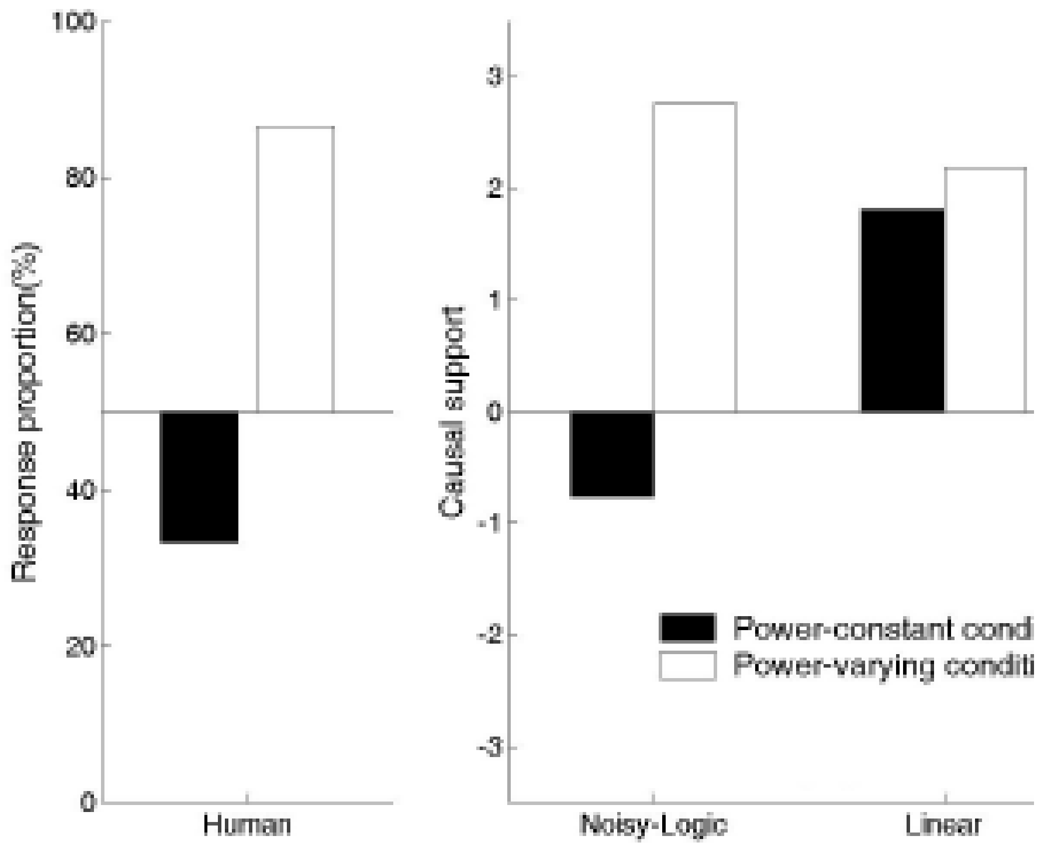
Based on the results from ALL THREE experiments, do you think that Medicine A interacts with some factor that varies across experiments, or do you think that the medicine influences the patients in different experiments in the same way?

Causal interaction

- Model selection



Causal interaction



Summary

- Bayes provides a *framework* for psychological theories
- Knowledge consists of *priors* and *generating models* (likelihoods)
- Psychological theories specify priors and generative models
- Bayesian framework facilitates testing alternative psychological theories