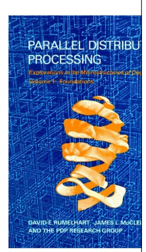
Bayes Days

cogs 105: computational modeling II: bayesian models

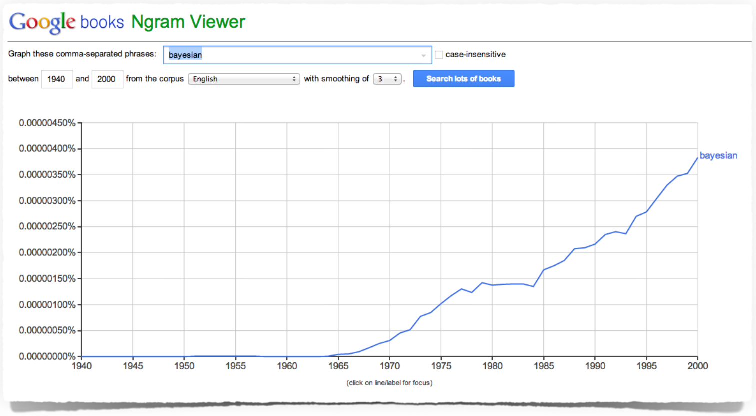
Must-Know Bayes

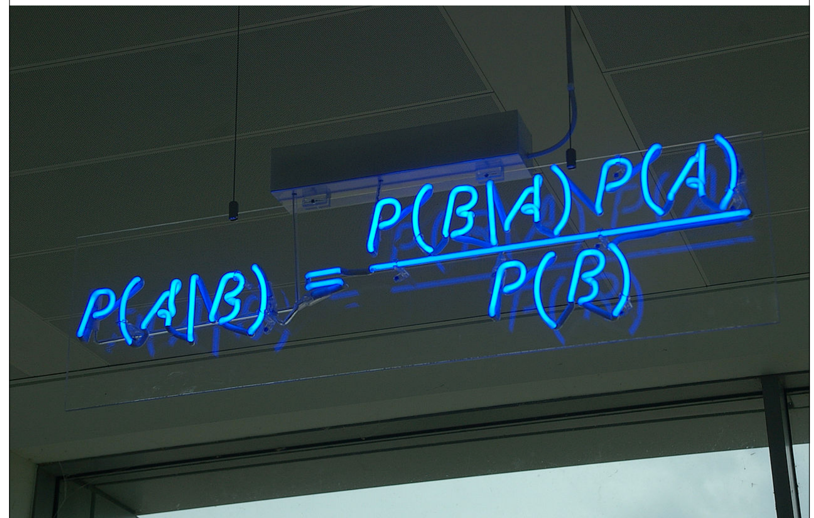
- Bayesian accounts of cognition are still on the rise; have become *heavily* influential in: perception, decision-making and reasoning, linguistics, categorization and category learning, semantic memory, …
- Rarely critiqued, until recently. E.g., *intense* critique: Marcus & Davis, 2013, in reading 2.

Bayes can be seen as a "mixed theory": it is both **symbolic**, but also intrinsically **probabilistic**



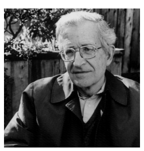
Should the human mind be seen as an engine of probabilistic inference, yielding *optimal* or *near-optimal* performance, as several recent prominent articles have suggested (Frank & Goodman, 2012; Gopnik, 2012; Téglás et al., 2011; Tenenbaum, Kemp, Griffiths, & Goodman, 2011)? Tenenbaum et al. (2011) argued that "over the past decade, many aspects of higher-level cognition have been illuminated by the mathematics of Bayesian statistics" (pp. 1279–1280), pointing to treatments of language; memory; sensorimotor systems; judgments of causal strength; diagnostic and conditional reasoning; human notions of similarity, representativeness, and randomness; and predictions about the future of everyday events.





$P(person \mid gender)$

Chomsky



$P(Chomsky)$

$P(male \mid Chomsky)$

Griffin

$P(male)$

Smith

$P(Chomsky \mid male)$

Bates

Varela

Bayes Theorem

H = Chomsky
D = male

$$P(H \mid D) = \frac{P(D \mid H)\ P(H)}{P(D)}$$

Bayes Theorem

H = Chomsky
D = male

$$P(H \mid D) = \frac{1.0 \;\; x \;\; 0.2}{0.4}$$

Bayes Theorem

H = Chomsky
D = male

$$P(H \mid D) = \frac{1.0 \;\; x \;\; 0.2}{0.4} = \mathbf{0.5}$$

One Way to View It

the probability the random person is Chomsky given that person is male…

H = Chomsky
D = male

$$P(H \mid D)$$

…proportional to…

$$P(D \mid H)\; P(H)$$

…the probability that you'd see a male given it is Chomsky…

…weighted by how probable Chomsky is in the first place!

One Way to View It

the probability there is rain given that it's January…

H = rain
D = January

$$P(H \mid D)$$

…proportional to…

$$P(D \mid H)\; P(H)$$

…the probability that it is January given that it's raining…

…weighted by how probable raining is in the first place!

In practice it's a bit more complicated…

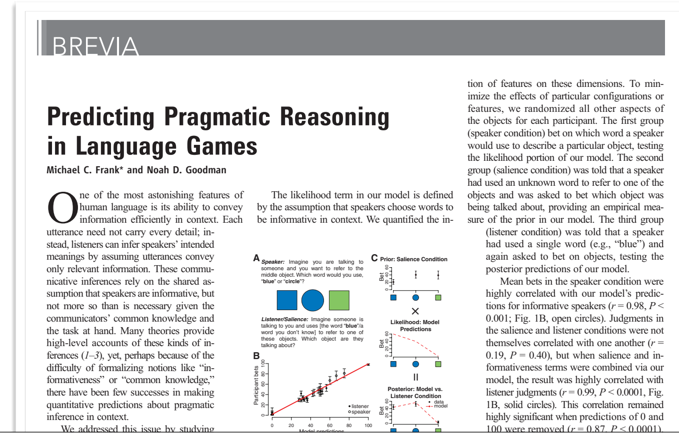
$$p(\mathrm{H}_i \mid D) = \frac{p(D \mid \mathrm{H}_i) \cdot p(\mathrm{H}_i)}{\sum_{j=1}^{k} p(D \mid \mathrm{H}_j) \cdot p(\mathrm{H}_j)}$$

In this equation, $p(\mathrm{H}_i \mid D)$ is the *posterior* probability of the hypothesis H_i given that the data D have been observed; $p(\mathrm{H}_i)$ is the *prior* probability that H_i is true before any data have been observed; and $p(D \mid \mathrm{H}_i)$ is the *likelihood*, the conditional probability that D would be observed assuming that H_i is true. The formula states that the posterior probability is proportional to the product of the prior probability and the likelihood. In most of the

From Marcus & Davis reading

First, an Example

Beautiful paper by Frank & Goodman, 2012



BREVIA

Predicting Pragmatic Reasoning in Language Games

Michael C. Frank* and Noah D. Goodman

One of the most astonishing features of human language is its ability to convey information efficiently in context. Each utterance need not carry every detail; instead, listeners can infer speakers' intended meanings by assuming utterances convey only relevant information. These communicative inferences rely on the shared assumption that speakers are informative, but not more so than is necessary given the communicators' common knowledge and the task at hand. Many theories provide high-level accounts of these kinds of inferences (*1–3*), yet, perhaps because of the difficulty of formalizing notions like "informativeness" or "common knowledge," there have been few successes in making quantitative predictions about pragmatic inference in context.

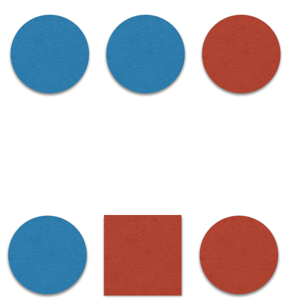
We addressed this issue by studying

The likelihood term in our model is defined by the assumption that speakers choose words to be informative in context. We quantified the in-

tion of features on these dimensions. To minimize the effects of particular configurations or features, we randomized all other aspects of the objects for each participant. The first group (speaker condition) bet on which word a speaker would use to describe a particular object, testing the likelihood portion of our model. The second group (salience condition) was told that a speaker had used an unknown word to refer to one of the objects and was asked to bet which object was being talked about, providing an empirical measure of the prior in our model. The third group (listener condition) was told that a speaker had used a single word (e.g., "blue") and again asked to bet on objects, testing the posterior predictions of our model.

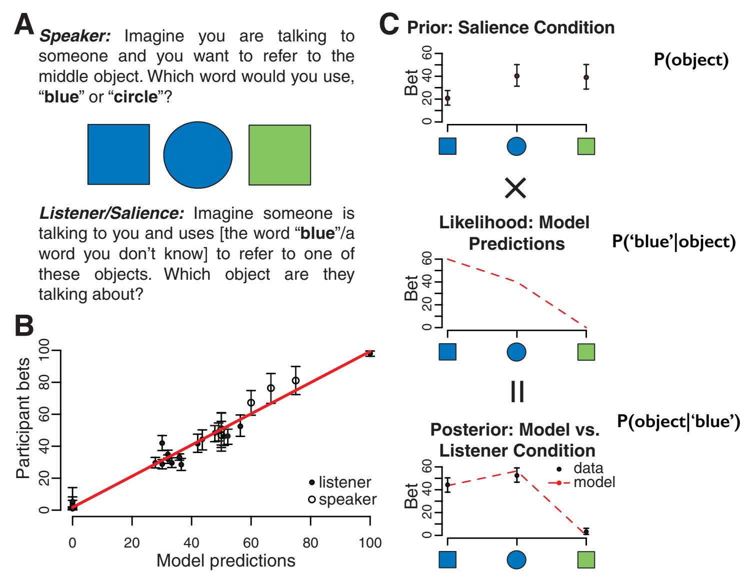
Mean bets in the speaker condition were highly correlated with our model's predictions for informative speakers ($r = 0.98$, $P < 0.001$; Fig. 1B, open circles). Judgments in the salience and listener conditions were not themselves correlated with one another ($r = 0.19$, $P = 0.40$), but when salience and informativeness terms were combined via our model, the result was highly correlated with listener judgments ($r = 0.99$, $P < 0.0001$, Fig. 1B, solid circles). This correlation remained highly significant when predictions of 0 and 100 were removed ($r = 0.87$, $P < 0.0001$).

"Pragmatic Inference"

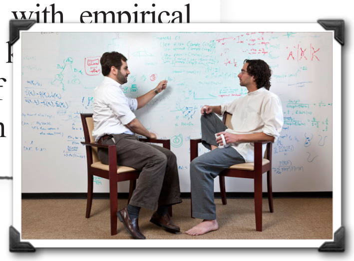


? *the red one*

? *the red one*



Our simple model synthesizes and extends work on human communication from a number of different traditions, including early disambiguation models (*8*), game-theoretic signaling models (*9*), and systems for generating referring expressions (*10*). The combination of an information-theoretic definition of "informativeness" along with empirical measurements of common [...] ables us to capture some of [...] human pragmatic inference in [...]



bark).

Couching their theory in the language of evolution and adaptation, Tenenbaum et al. (2011) argued that "the Bayesian approach [offers] a framework for understanding why the mind works the way it does, in terms of rational inference adapted to the structure of real-world environments" (p. 1285).

To date, these models have been criticized only rarely (Bowers & D[...]

subjects' responses to seven of the nine questions. Griffiths and Tenenbaum concluded that "everyday cognitive judgments follow the . . . optimal statistical principles" and there is "close correspondence between

"RATIONAL"

"OPTIMAL"

General Article

Psychological Science
XX(X) 1–10

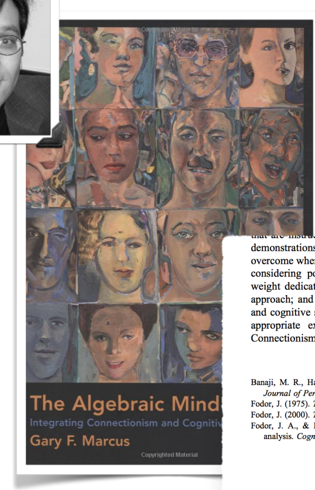
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797613495418
pss.sagepub.com
SAGE

How Robust Are Probabilistic Models of Higher-Level Cognition?

Gary F. Marcus and Ernest Davis
New York University

Abstract
An increasingly popular theory holds that the mind should be viewed as a *near-optimal* or *rational* engine of probabilistic inference, in domains as diverse as word learning, pragmatics, naive physics, and predictions of the future. We argue that this view, often identified with Bayesian models of inference, is markedly less promising than widely believed, and is undermined by post hoc practices that merit wholesale reevaluation. We also show that the common equation between *probabilistic* and *rational* or *optimal* is not justified.

Keywords
cognition(s), Bayesian models, optimality



demonstrations of network failings, he offers lessons on what to avoid and overcome when designing connectionist models. The book does suffer from only considering potential problems of eliminative connectionism, without equal weight dedicated to discussing the well-known shortcomings of the symbolic approach; and this despite a subtitle that promises to *integrate* connectionism and cognitive science. A more suitable subtitle, one that would at least lead to appropriate expectations for the reader, might have been ''Assimilating Connectionism into Symbolic Cognitive Science''.

REFERENCES

Banaji, M. R., Hardin, C., & Rothman, A. J. (1993). Implicit stereotyping in person judgment. *Journal of Personality and Social Psychology*, *65*, 272–281.
Fodor, J. (1975). *The language of thought*. New York: Crowell.
Fodor, J. (2000). *The mind doesn't work that way*. Cambridge, MA: MIT Press.
Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *78*, 3–71.

RICK DALE AND MORTEN H. CHRISTIANSEN
Cornell University, Ithaca, USA

Two Big Problems

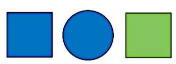
- Task selection
  - Simple: how do we know what does and doesn't work? "Breadth first" vs. "depth first" strategy
- Model selection
  - There is not one unique model for any given task or situation; how is this selected?

Table 1. Examples of Domains in Which Performance Has Been Found to Fit Naturally With Probabilistic Explanations in Some Cases but Not Others

Domain	Apparently optimal performance	Apparently nonoptimal performance
Intuitive physics	Tower problems (Battaglia, Hamrick, & Tenenbaum, in press)	Balance-scale problems (Siegler, 1976) Projectile-trajectory problems (Caramazza, McCloskey, & Green, 1981)
Incorporation of base rates	Various tasks (Frank & Goodman, 2012; Griffiths & Tenenbaum, 2006)	Base-rate neglect (Kahneman & Tversky, 1973; but see Gigerenzer & Hoffrage, 1995)
Extrapolation from small samples	Future prediction (Griffiths & Tenenbaum, 2006) Size principle (Tenenbaum & Griffiths, 2001a)	Anchoring (Tversky & Kahneman, 1974) Underfitting of exponentials (Timmers & Wagenaar, 1977) Gambler's fallacy (Tversky & Kahneman, 1974) Conjunction fallacy (Tversky & Kahneman, 1983) Estimating unique events (Khemlani, Lotstein, & Johnson-Laird, 2012)
Word learning	Using sample diversity as a cue to induction (Xu & Tenenbaum, 2007)	Using sample diversity as a cue to induction (Gutheil & Gelman, 1997) Evidence selection (Ramarajan, Vohnoutka, Kalish, & Rhodes, 2012)
Social cognition	Pragmatic reasoning (Frank & Goodman, 2012)	Attributional biases (Ross, 1977) Egocentrism (Leary & Forsyth, 1987) Behavioral prediction of children (Boseovski & Lee, 2006)
Memory	Rational analysis (Anderson & Schooler, 1991)	Eyewitness testimony (Loftus, 1996) Vulnerability to interference (Wickens, Born, & Allen, 1963)
Foraging	Animal behavior (McNamara, Green, & Olsson, 2006) Information foraging (Jacobs & Kruschke, 2011)	Probability matching (West & Stanovich, 2003)
Deductive reasoning	Deduction (Oaksford & Chater, 2009)	Deduction (Evans, 1989)
Overview	Higher-level cognition (Tenenbaum, Kemp, Griffiths, & Goodman, 2011)	Higher-level cognition (Kahneman, 2003; Marcus, 2008)

$$P(r)$$

Salience task: "Imagine someone uses a word you don't know; which do you think they'd refer to?



Surprisal model for likelihood

$$P(w \,/\, r, C)$$

$|w|$ # of objects word w can be used in.

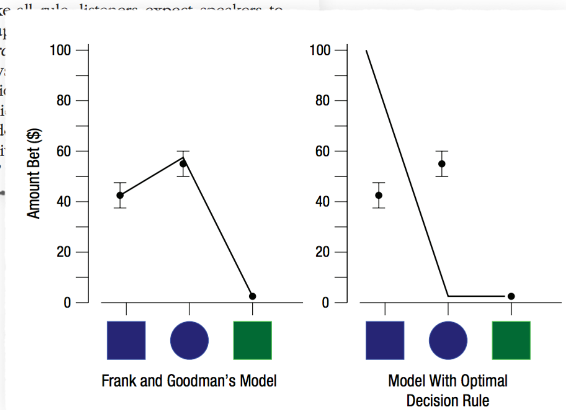
W # of words that could be used for the intended object.

$$P(w|r_s, C) = \frac{|w|^{-1}}{\sum_{w' \in W} |w'|^{-1}}$$

would use *circle* two thirds of the time and *blue* one third of the time. Although this decision rule is not uncommon, Frank and Goodman might just as easily have chosen a model with a winner-take-all decision rule, following the maximum-expected-utility principle, which is the standard rule in decision theory. According to the winner-take-all rule, listeners expect speakers to always use the applicable word of greatest specificity; this would be *circle* if the middle object were intended. As Figure 4 shows, although the model with Frank and Goodman's decision rule yielded a good fit to the data, other models, which are actually more justifiable a priori, would have yielded dramatically poorer fits. Details of the analysis are given in the Supplemental Material.

Experimenters' choice of how to word the questions posed to subjects can also affect model fit. For example,

would use *circle* two thirds of the time and *blue* one third of the time. Although this decision rule is not uncommon, Frank and Goodman might just as easily have chosen a model with a winner-take-all decision rule, following the maximum-expected-utility principle, which is the standard rule in decision theory. According to the winner-take
always use the a
this would be *circ*
As Figure 4 show
Goodman's decisi
other models, whi
would have yield
the analysis are gi
Experimenters'
posed to subjects



acter but incorrect in their assumptions.)

More broadly, probabilistic models have not yielded a robust account of cognition. They have not converged on a uniform architecture that is applied across tasks; rather, there is a family of different models, each depending on highly idiosyncratic assumptions tailored to an individual task. Whether or not the models can be said to fit depends on the choice of task, how decision rules are chosen, and a range of other factors. The Bayesian approach is by no means unique in being vulnerable to these criticisms, but at the same time, it cannot be considered to be a fully developed theory until these issues are addressed.

The greatest risk, we believe, is that probabilistic

Check Their Response

Relevant and robust. A response to Marcus and Davis.

Noah D. Goodman

Department of Psychology, Stanford University

Michael C. Frank

Department of Psychology, Stanford University

Thomas L. Griffiths

Department of Psychology, University of California, Berkeley

Joshua B. Tenenbaum

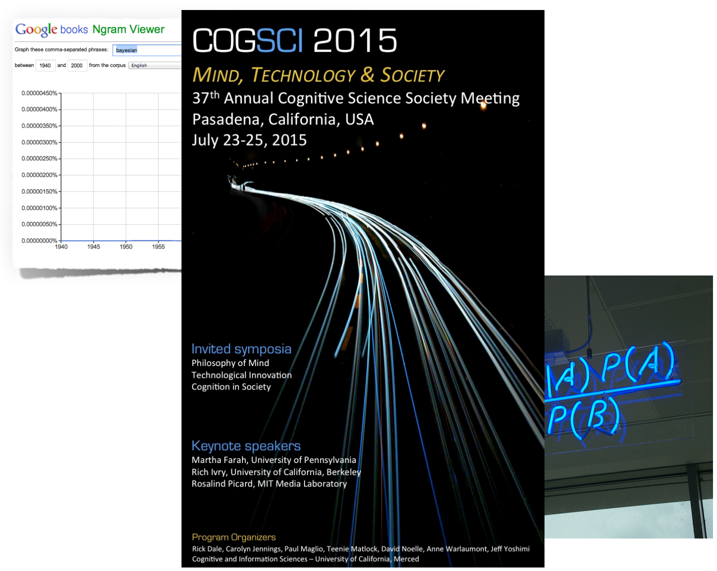
Department of Brain and Cognitive Sciences, MIT

Peter Battaglia

Department of Brain and Cognitive Sciences, MIT

Jessica Hamrick

Department of Psychology, University of California, Berkeley



Next Week

- What to do?
- Final paper
- Applied issues… careers!