

The Role of Causal Models in Statistical Reasoning

Tevya R. Krynski (tevyar@mit.edu)

Joshua B. Tenenbaum (jbt@mit.edu)

Department of Brain & Cognitive Sciences, Massachusetts Institute of Technology
77 Massachusetts Ave, Cambridge, MA 02139 USA

Abstract

When making judgments based on statistical data, people have been shown to be poor at probabilistic reasoning, specifically Bayesian inference. While recent research shows people perform better when information is provided in a natural frequency format, we find this result's explanatory reach limited, both for explaining people's judgment failures and as a theory of human reasoning under uncertainty. Most prior studies demonstrating probabilistic reasoning deficits gave their subjects probabilistic inference problems that did not explain the causal mechanisms behind the provided statistics. Our research shows that when questions are posed that explain the causal structure of the domain, subjects perform significantly better. Specifically, base rate neglect can be made to virtually disappear when the content of a question reflects the true causal structure of the domain. We propose that causality is essential to probabilistic reasoning, and that without the opportunity to incorporate statistical data into a consistent theory of the causal structure of a domain, the typical person will have trouble performing normatively correct Bayesian inference.

Introduction

Can people arrive at normatively correct probability judgments after reading sufficient statistical data? A decades-old subject of experimental inquiry, statistical inference has proven to be problematic for most people to grasp. Examples include the well-studied phenomena of base-rate neglect (Kahneman & Tversky, 1982; Bar-Hillel, 1980), the conjunction fallacy (Tversky & Kahneman, 1983), and deviations from the additivity principle (Villejoubert & Mandel, 2002). Yet people clearly function extremely well in the everyday world, an environment saturated with useful statistical information. Could it be that people are consistently forming incorrect judgments and making wrong decisions in the face of statistical data that would provide the correct answer, if only they could compute it? This surely happens occasionally, but thus far cognitive science has not accounted for why people reason so well so often. It is possible that we have not yet begun to test the true capability of human statistical competence.

Bayesian inference and base-rate neglect

This paper focuses on base-rate neglect, the phenomenon first named by Kahneman and Tversky for the case when people seemingly neglect the base rate of a hypothesis while performing probabilistic inference. The widely accepted normative method of performing probabilistic inference is to

use Bayesian inference, which requires the following four elements:

1. Hypothesis (H): the statement you wish to judge the probability of.
2. Data (D): an observation that has been made that can be used as evidence for or against the hypothesis.
3. Prior $P(H)$: your belief that the hypothesis was true before making the observation.
4. Likelihoods $P(D|H)$ and $P(D|\neg H)$: the probabilities that the observation would have been observed if the hypothesis were true and if the hypothesis were false.

Given these, Bayes' theorem prescribes an equation for computing the "posterior" (the probability of the hypothesis given the data): $P(H|D) = P(H) \times P(D|H) / P(D)$, where $P(D)$ comes from $P(H) \times P(D|H) + P(\neg H) \times P(D|\neg H)$. Bayes' theorem does not prescribe how one should set the prior probability or the likelihoods, but most researchers have assumed that subjects should set them based on the statistics provided in the experiment.

The primary example of base-rate neglect that this paper considers is a word problem adapted from Eddy (1982) and tested in an influential paper by Gigerenzer and Hoffrage (1995). This problem forms the basis of a discussion continued in numerous subsequent papers and studies (Macchi, 2000, Lewis & Keren, 1999, Cosmides & Tooby, 1996), and is the primary example used in Gigerenzer's influential natural frequency theory. We therefore believe it is worth subjecting it to a detailed analysis. The problem reads as follows (from Gigerenzer & Hoffrage, 1995):

The probability of breast cancer is 1% for a woman at age forty who participates in a routine screening. If a woman has breast cancer, the probability is 80% that she will get a positive mammography. If a woman does not have breast cancer, the probability is 9.6% that she will also get a positive mammography. A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer? ____ %

Subjects are poor at solving this problem, often giving an answer of 70%-90%, while Bayes' theorem prescribes an answer of 7.8% (Gigerenzer & Hoffrage, 1995). Kahneman & Tversky (1982) used the term "base-rate neglect" to characterize errors in this range because in this and similar problems subjects seemed to be neglecting the base rate of 1% (the prior) in favor of the individuating information (the likelihoods), when in fact they should be combining the two to calculate the posterior. While some of the answers subjects give are consistent with the theory that they neglect the base rate, other explanations have been offered, such as the tendency to confuse a given conditional probability with its inverse (the inverse fallacy) (Villejoubert & Mandel, 2002). Laura Macchi (1995) has catalogued numerous incorrect answers to typical inference problems, and while many of them fall into the range characterized by base rate

neglect, few subjects actually neglect the base rate by calculating $P(H|D) = P(D|H) / [P(D|H) + P(D|\neg H)]$. More often they calculate $P(H|D)$, $1 - P(D|\neg H)$, or $P(D|H) - P(D|\neg H)$.

Regardless of how one categorizes errors, one thing is clear: people do not possess a general-purpose probabilistic reasoning engine that takes statistical data as input and produces correct probabilities as output. Gigerenzer's influential natural frequency research program has shown that questions provided in a natural frequency format, rather than a probabilistic format, can make inference errors such as base rate neglect disappear. As an example, Gigerenzer & Hoffrage (1995) demonstrated that people perform much better when answering the following version of the mammography problem in a "natural frequency format":

- 10 out of every 1,000 women at age forty who participate in a routine screening have breast cancer.
 - 8 of every 10 women with breast cancer will get a positive mammography.
 - 95 out of every 990 women without breast cancer will also get a positive mammography.
- Here is a new representative sample of women at age forty who got a positive mammography in a routine screening. How many of these women do you expect to actually have breast cancer? ___ out of ___.

Natural frequency proponents have concluded that people have no cognitive algorithm for working with information in probability format: "We assume that as humans evolved, the 'natural' format was frequencies as opposed to probabilities or percentages... The evolutionary argument that cognitive algorithms were designed for frequency information, acquired through natural sampling, has implications for the computations an organism needs to perform when making Bayesian inferences" (Gigerenzer & Hoffrage, 1995).

While it is well-established that people perform better on frequency formats, we find this result limited as an explanation for probabilistic reasoning errors, because it does not address how people compute probabilities, nor why they make mistakes. The conclusion that people cannot make use of probabilistic information ignores a substantial body of research showing that people perform well on certain classes of probability questions, particularly those in which the base rate has been variously called "causal" (Kahneman & Tversky, 1982), "salient", "relevant", and "specific" (Bar-Hillel, 1980). The issue of how they manage this, given their supposed lack of a probability engine, has been unaddressed and unanswered by frequency format researchers. Instead, Gigerenzer & Hoffrage seek to "shift the focus from human errors to human engineering: how to help people reason the Bayesian way without even teaching them." (Gigerenzer & Hoffrage, 1995).

Although it is clear subjects perform better on frequency formats, the frequentist approach may not be such a substantive test of Bayesian inference. Specifically, the question being asked of people in the natural frequency format is not really a problem of inference, but instead a problem of proportions. Bayesian inference problems, as the frequentist proponents themselves define it, are "problems in which the probability of a cause (e.g. cancer) has to be inferred from an observed effect (e.g. a positive mammography result)." (Hoffrage, et al, 2002) However, the frequency format on which they base their theory does

not ask for a probability of a cause nor does it provide an observed effect. In Gigerenzer & Hoffrage (1995), the question being asked is:

Here is a new representative sample of women at age forty who got a positive mammography in a routine screening. How many of these women do you expect to actually have breast cancer? ___ out of ___.

The authors nevertheless insist that subjects use Bayesian inference to solve this problem. We believe the authors are asking for a proportion, and that is what they are getting: the proportion of positive-testing women that have cancer. To answer this question, one must count the total number of women who test positive (8 plus 95), and count the number of them who have cancer (8). Nowhere in this calculation is the probability of a cause inferred from an observed effect, nor is there any need to think causally.

Suppose, for the sake of argument, that rather than asking for a proportion, the authors of the frequency format were to ask for the probability of cancer given a positive mammogram. Although we do not know if this has ever been tested, let us grant, for the moment, that people would do well on this problem, giving a correct answer of 8/103 much of the time. Would this qualify as Bayesian inference? Not necessarily. There are actually two commonly used formulas for conditional probability. In addition to Bayes' rule, $P(H|D) = P(H) \times P(D|H) / P(D)$, there is also the traditional definition of conditional probability: $P(H|D) = P(H \& D) / P(D)$ (note that these are equivalent because $P(H) \times P(D|H) = P(H \& D)$). For a large enough number of samples, the traditional definition of conditional probability is approximately equal to the conditional relative frequency: $N(H \& D) / N(D)$, where $N(X)$ is the number of occurrences of X . It is likely that most subjects would use this relative frequency formula to calculate the probability using the data provided in the frequency format, and would still not be using Bayes' rule.

By asserting that computing proportions (or conditional relative frequencies) qualifies as Bayesian inference, we believe the frequentist approach bypasses the most important issues of reasoning under uncertainty. Specifically, it does not address the question Kahneman & Tversky were interested in: when and why people correctly integrate prior probabilities with new evidence. Furthermore, it reduces inference to the calculation of proportions, which could not possibly account for the unexplained human capability to reason soundly without complete statistical data. Judgments such as the probability of a football team winning, the chance of getting a certain job, or a con artist's intentions, are all usually made without reference to a large collection of identical past experiences required to compute frequentist proportions.

Bayesian inference in a causal framework

We propose that rather than possessing a domain-general probabilistic engine taking statistical data as input and producing probabilities of hypotheses as output, people must evaluate and interpret the statistical information that does exist within the framework of a probabilistic causal model: a theory of how causes produce effects. An individual's probabilistic causal model encompasses

knowledge of which causes produce which effects (the structure), how likely certain causes are (the priors) and how likely a given effect is to follow from a given set of causes (the likelihoods). This model provides the knowledge base for a hypothesized causal reasoning engine, which takes as input (1) a probabilistic causal model and (2) observations or data, and is capable of producing probabilities of hypotheses as output. The causal reasoning engine can be formally modeled as a Bayesian network, a graphical computational model used to represent causal influence and to perform probabilistic inference (Pearl, 2000).

Recently, researchers have used Bayesian networks for modeling several domains of human reasoning, including category learning (Rehder, 2001), inferring causality from covariation (Ahn & Kalish, 2000), causal chain induction (Ahn & Dennis, 2000), causal learning in children (Gopnik et al, in press), causal structure learning (Tenenbaum & Griffiths, 2001) and causal inference from observations and interventions (Steyvers et al., 2002). We believe Bayesian networks mark a significant advance in understanding human reasoning under uncertainty, as they enable us to propose, model, and test formal theories of causal reasoning.

We intend to argue that the difficulty in the probabilistic version of the mammogram problem stems not from neglecting the base rate, but from misunderstanding the causal mechanism behind the false positive rate. Kahneman & Tversky (1982) proposed that base rates with a causal explanation improved inference; we expand on that idea by being more specific about how causal knowledge maps onto the components of a causal model. We maintain that one primary condition for correct reasoning is that the question fit with the subject's prior causal model of the domain in question. If the descriptions of the data are arbitrary, deceptive, or otherwise unrealistic, subjects may have a difficult time, particularly if their prior knowledge of the domain is contradicted by the data provided. In this case, the statistics provided in the mammogram problem contain significant deviations from the true statistics.

Gigerenzer and Hoffrage adapted the probabilistic version of the breast cancer problem from David Eddy's 1982 article describing the statistical nature of mammograms. Looking at the original Eddy (1982) paper, we found a number of important discrepancies between it and Gigerenzer and Hoffrage's adaptation that could be at least partly responsible for subjects' poor performance.

1. In Eddy's paper, the cancer rate of 1% was not for women receiving a routine screening; in fact, it wasn't even a real statistic. It was just an example of one physician's "subjective" ("perhaps subconscious") estimate of the frequency of cancer among women, who are "in all important respects... are similar" to "a patient *with a breast mass*" (italics added) (Eddy, 1982)
2. In Eddy's paper, the likelihoods of 80% and 9.6% are also not for women receiving routine screenings. The numbers come from Snyder (1966, p. 217), whose statistics are of women *who already have a breast mass (a lesion)*: "The results showed 79.2 per cent of 475 malignant lesions were correctly diagnosed and 90.4 per cent of 1,105 benign lesions were correctly diagnosed"

(Snyder, 1966). Gigerenzer and Hoffrage erroneously applied the likelihood of 9.6% to all women without cancer, which means subjects have no indication that the lesions are actually the cause of the false positives.

3. Even if the above discrepancies were fixed, the structure of the problem is misleading, by giving a probability of 9.6% that a woman without cancer will get a positive mammography. This could be interpreted to mean that if this woman takes the mammogram 1000 times, she will receive a positive result approximately 96 times. However, the facts of the matter are quite different: the size and density of the benign lesion is actually the major determinant of the false positive, and this does not change from moment to moment. So, while it is true that 9.6% of women with benign lesions will receive a positive mammogram, it is not true that any individual will have a 9.6% chance; some will have a high chance and others a low chance.

Combining the above 3 discrepancies, one gets the impression from the given problem that the mammogram is an extremely error-prone test: due to random error, the mammogram will come back positive nearly 10% of the time when testing a woman without cancer, for no reason whatsoever. How could the medical community trust such an error-prone test? Why would anyone use it? These discrepancies are crucial; we believe the reason subjects perform so poorly on this problem is that they disregard or under-weight the false positive rate due to their prior knowledge that the doctors would not trust such a wildly error-prone test.

Probabilistic Causal Models

A basic causal model for this scenario is depicted in Figure 1, in which the probability of a positive mammogram depends on whether or not the patient has cancer. This is only the simplest possible model; others are also compatible with the statistics. Suppose a subject believes that a positive mammogram is tantamount to a doctor's diagnosis of breast cancer. This is not implausible: doctors have a duty to avoid scaring their patients unnecessarily, and it is common for a doctor to say: "you have some indications consistent with disease X, but it's probably nothing". So, if the doctor says: "you've tested positive for breast cancer," rather than "the preliminary tests indicate a possibility of cancer, but it's probably nothing", there is a good chance you have cancer. In this case, the subject's model might look more like the one depicted in Figure 2, in which the patient's having cancer causes both a positive mammogram and other evidence that informs the doctor's diagnosis of cancer.

To use the structure of Figure 2 as a Bayesian network, one must supply the probability of a cancer diagnosis given that a patient has (or doesn't have) evidence of cancer. Suppose the subject estimates that a healthy person has a 1% chance of being told they have tested positive for cancer, and a person with cancer has a 50% chance of being told they have cancer (as opposed to being told that the results are inconclusive and more tests have to be done). Using Bayes' rule, the normatively correct answer for the probability of cancer given a positive result is 80.8% (or

$1\% \times 80\% \times 50\% / [1\% \times 80\% \times 50\% + 99\% \times 9.6\% \times 1\%]$. While this calculation is consistent with answers subjects often provide, we are not claiming that this is the actual model they use when they commit base rate neglect. We are simply demonstrating that one source of “error” could be subjects’ correct use of prior domain knowledge.

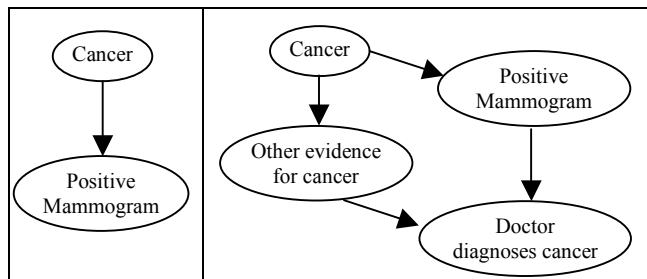


Figure 1: basic causal model

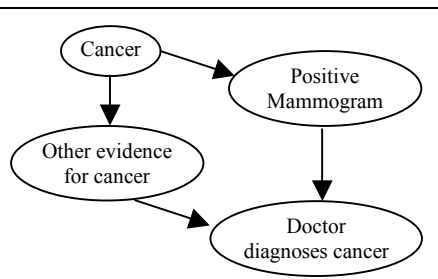


Figure 2: causal model including doctor’s diagnosis

The problem with both the models of Figures 1 and 2 is that they do not reflect the true causal structure of the test. Figure 3 fits Snyder’s statistics more accurately. In this model, the source of the false positives is clear: dense benign lesions. Specifically, 9.6% of the women without cancer (but with a breast mass) have a benign lesion dense enough to cause a positive mammogram.

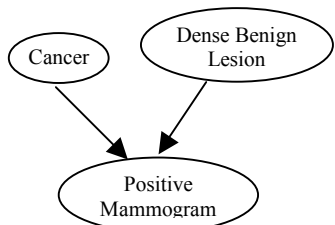


Figure 3: Simple causal model to explain false positives

Experiment 1

Our first experiment directly tested the idea that subjects may be tacitly assuming the positive mammogram result to be tantamount to a doctor’s diagnosis of cancer. We hypothesized that subjects would perform better if most women without cancer received an “uncertain” mammogram result rather than a “positive” one.

Method

Participants Our subjects were airplane passengers, approached while waiting for their flights to begin boarding. Their only compensation was the temporary alleviation of boredom through engaging their mind with our question.

Materials We gave our subjects paper and pen versions of Gigerenzer’s classic breast cancer question, except that the mammogram had three possible results: positive, uncertain, and negative (some mammograms do have three possible results, according to Eddy, 1982). Subjects got one of two versions of this question: one in which a woman receives a positive result during a routine screening, and one in which

she receives an uncertain result. The numbers were exactly the same in both versions, except that the conditional probabilities for positive and uncertain were switched, so that the subjects were required to perform the same calculation in both versions. The questions follow:

“Positive” Question

From past statistics of routine mammography screenings, the following is known:

1% of the women who have participated in past screenings had breast cancer at the time of the screening.

Of the 1% who had breast cancer, 20% tested ‘uncertain’ during the mammogram (further testing was required to determine that they had breast cancer), and the other 80% tested ‘positive’.

Of the 99% of women who did not have breast cancer, 2% tested ‘uncertain’ (further testing was required to determine that they did not have breast cancer), 9.6% tested ‘positive’, and the other 88.4% tested ‘negative’.

Suppose a woman in this age group participates in a routine mammography screening and the test result is ‘positive’. Without knowing any other symptoms, what is the probability that she actually has breast cancer?

“Uncertain” Question

From past statistics of routine mammography screenings, the following is known:

1% of the women who have participated in past screenings had breast cancer at the time of the screening.

Of the 1% who had breast cancer, 20% tested ‘positive’ during the mammogram, and the other 80% tested ‘uncertain’ (further testing was required to determine that they had breast cancer).

Of the 99% of women who did not have breast cancer, 2% tested ‘positive’, 9.6% tested ‘uncertain’ (further testing was required to determine that they did not have breast cancer), and the other 88.4% tested ‘negative’.

Suppose a woman in this age group participates in a routine mammography screening and the test result is ‘uncertain’. Without knowing any other symptoms, what is the probability that she actually has breast cancer?

Results

A one-way ANOVA of the raw responses revealed a significant difference between the two versions ($F(1,71)=21.59$, $MSE=897.36$, $p<.0001$). The results strongly suggest that subjects perceive a positive test result to be more diagnostic of cancer than an uncertain test result, even though the calculation required is identical. We classified base rate neglect as any answer greater than or equal to 70%. Since the actual correct answer of 7.8 is difficult to calculate from the problem (only 2 subjects got it), we classified as “close” any answer between 5% and 12% inclusive. The result was a significant difference in base rate neglect between the two versions ($\chi^2(2) = 12.49$, $p < .0005$).

Table 1: Positive vs. Uncertain, Airport Passengers

Mammogram	Base Rate Neglect	Close Answer	Neither
Positive	14	9	12
Uncertain	1	15	22

Experiment 2

Our second experiment tested our hypothesis that subjects do not understand the causal mechanism behind the false positive, which results in them under-weighting the false positive rate. We hypothesized that subjects would perform better on the mammogram question when informed that

dense benign lesions are the cause of the false positives, as depicted in Figure 2.

Method

Participants For this experiment, we used both airport passengers and MIT students, compensating the MIT students with candy, and airport passengers with the temporary alleviation of boredom.

Materials We posed two paper and pen questions, one in which there was only “probabilistic” information about false positives, and one in which there was “causal” information. Crucially, both versions required the exact same Bayesian formula to calculate the answer. We used a within-subjects design, which required two questions per subject. In order to provide two seemingly different questions, we varied four factors for the second question: the base rate (1% or 2%), the false positive likelihood (4% or 6%), the true positive likelihood (80% and 90%) and the cover story (breast cancer vs. harmless cyst or colon cancer vs. harmless polyp). Half the subjects received the “probabilistic” version followed by the “causal”, and the other half received the “causal” followed by the “probabilistic”. The numbers, cover story, and causal information were counterbalanced for a total of 8 different packets. To minimize arithmetic errors, the subjects were allowed to answer as either a ratio or a percentage. The questions follow¹:

“Probabilistic” Question

The following statistics are known about women at age 60 who participate in a routine mammogram screening, an X-ray of the breast tissue that detects tumors:

About 2% have breast cancer at the time of the screening. Most of those with breast cancer will receive a positive mammogram.

There is about a 6% chance that a woman without cancer will receive a positive mammogram.

Suppose a woman at age 60 participates in a routine mammogram screening and receives a positive mammogram. Please estimate the chance that she actually has breast cancer.

“Causal” Question

The following statistics are known about women at age 60 who participate in a routine mammogram screening, an X-ray of the breast tissue that detects tumors:

About 2% have breast cancer at the time of the screening. Most of those with breast cancer will receive a positive mammogram.

About 6% of those without cancer have a dense but harmless cyst, which looks like a cancerous tumor on the X-ray and thereby results in a positive mammogram.

Suppose a woman at age 60 participates in a routine mammogram screening and receives a positive mammogram. Please estimate the chance that she actually has breast cancer.

Note that this experiment did not include the true positive rate, but instead stated, “Most women with breast cancer will receive a positive mammogram.”² This was done to

¹ Full questions and instructions for both experiments can be found at <http://web.mit.edu/tevya/www/CogSci2003>

² A third study produced a significant base rate neglect difference between the original problem, with the true positive rate of 80%, and a problem that replaced the true positive rate with the rate of benign cysts, but did not state that benign cysts caused positive mammograms (see <http://web.mit.edu/tevya/www/CogSci2003>).

encourage subjects to provide answers based on their intuition rather than memorized mathematical formulas. It had the added benefit of demonstrating that the true positive rate did not have to be given in order to induce base rate neglect, contradicting the hypothesis that the inverse fallacy is the primary cause of base rate neglect.

Results

Preliminary analyses showed no differences between MIT students and airport passengers, so the two groups were collapsed for the remaining analyses. A three-way ANOVA of the raw responses to the first question answered showed no significant interactions, with a significant difference between “Probabilistic” and “Causal” questions ($F=8.33$, $p<.005$), and no significant effect of cover story ($F=0.43$, $p=.51$) or prior/likelihood values ($F=.0052$, $p=.94$) all with $df=(1,125)$, $MSE=836$. We again classified base rate neglect as any answer greater than or equal to 70%. To obtain a correct answer in this version, the subject had to supply the correct ratio or percentage, or a number up to 20% below it (to accommodate the fact that most, but not all, women with cancer receive a positive result). The causal version significantly reduced base rate neglect and significantly improved correct answer rates as compared to the probabilistic version ($\chi^2(2) = 12.83$, $p < .0005$) (see table 2).

Table 2: Probabilistic vs. Causal

Problem Type	Base Rate Neglect	Correct	Neither
Probabilistic	19	17	40
Causal	3	31	45

The within-subjects design allowed us to track how each subject performed on both versions. Our results show that subjects improved from 22% correct to 32% correct going from the probabilistic (P) to the causal (C) version. They also declined from 39% correct to 33% correct going from C to P. Similarly, base rate neglect declined from 24% to 20% going from P to C, and increased from 4% to 10% going from C to P. This demonstrates that, despite a strong carry-over effect, the causal version was easier within subjects, not just between subjects.

Discussion

By explaining the causal mechanism behind the false positive rate of the mammogram, we effectively eliminated the occurrence of base rate neglect. A total of 4 out of 117 subjects exhibited base rate neglect on our new questions, compared to 33 of 111 on the original version, despite the required calculations being identical. The first experiment indicates that base rate neglect in this case could derive from equating a positive mammogram to a diagnosis of cancer (the doctor would have labeled the result “uncertain” if he were not confident that the patient had cancer). Alternatively, the second experiment indicates that base rate neglect may reflect subjects’ inability to understand the mechanism by which a positive result could occur for a woman without breast cancer.

A simpler explanation, of course, is that subjects simply generally disregard the numbers, and just answer based on experience, picking an arithmetic combination of numbers

from the problem that comes closest to their intuitive feeling of what the answer should be. This explanation, however, raises the question: what does it mean to answer based on experience? Very few subjects have ever had a positive mammogram (especially not male MIT students). They cannot, therefore, know the answer to this question based on personal experience, nor from the experience of acquaintances, as people rarely discuss their mammogram results if they are found to be cancer-free through biopsy. But, subjects do have knowledge of doctors and medical care in general, and they know that a doctor is unlikely to tell a patient she has cancer if he believes the chances are low, or if the test is highly error-prone. This is exactly the sort of abstract knowledge that we believe subjects instantiate into their probabilistic causal models.

Base rate neglect has become a catchall phrase for a number of different phenomena. We saw many different errors, the most common of which were $P(C)$, $1 - P(Pos | \neg C)$, $P(Pos | C)$, $P(Pos | \neg C)$, 1 (100%), and a number of wild guesses (e.g., 50%, 60%). We note that the wide variety of answers given to these word problems makes them a crude measure of natural human reasoning competency. Many subjects, especially those selected from the general public, find word problems daunting and inform us that they simply ventured a wild guess. Equally harmful to our measurement are those subjects, especially university students, who apply formulas they learned in school without engaging their natural intuitive reasoning skills.

Even the frequentist trick of imagining 1000 patients, which we know many of our subjects used (from their handwritten calculations), is not the type of “sound” reasoning we wish to measure. This method is a domain-general strategy for use with mathematical word problems, yet people clearly do not use such strategies for real-world, complex systems. When is the last time you heard your mechanic say: “if you want to estimate the chances of your car breaking down on a long road trip, first imagine 1000 cars...” This trick only works when the statistics are completely known, and are given for a small number of variables. We believe future studies should be done in an engaging, interactive environment whose results are not affected by the variety of reasoning errors that can result from misreading a word, or misplacing a decimal point.

Our results show that people can more successfully perform Bayesian inference on probabilistic information if they are given a causal explanation for the data. We hypothesize that if a statistic cannot be accommodated in a subject’s causal model, it will be ignored or under-weighted. This could be due to conflicting prior knowledge, or lack of a suitable cause to explain the statistic. We believe that much of human reasoning under uncertainty employs probabilistic causal models, and that further inquiry into how these models are constructed and utilized will illuminate when and why people make sound inferences.

References

Ahn, W., & Kalish, C. (2000). The role of covariation vs. mechanism information in causal attribution. In R. Wilson, & F. Keil (Eds.) *Cognition and explanation*, Cambridge, MA: MIT Press.

- Ahn, W., & Dennis, M. (2000). Induction of causal chains. *Proceedings of the Twenty-second Annual Conference of the Cognitive Science Society*, Lawrence Erlbaum Associates, NJ: Mahwah
- Bar-Hillel, M. (1980) The base-rate fallacy in probability judgments. *Acta Psychologica*, 44 (pp. 211-233).
- Cosmides, L. and Tooby, J. (1996) Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, 58 (pp. 1-73)
- Eddy, David M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds), *Judgment under uncertainty: Heuristics and biases* (pp. 249-267). Cambridge, England: Cambridge University Press.
- Gigerenzer, G. & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological review*. Vol. 102, No. 4, 684-704
- Gigerenzer, G. (2000). *Adaptive Thinking*. New York: Oxford University Press.
- Gopnik, A., Glymour, C., Sobel D., Schulz L., Kushnir, T., & Danks, D. (in press). A theory of causal learning in children: Causal maps and Bayes-Nets. *Psychological Review*.
- Hoffrage, U., Gigerenzer, G., Krauss, S. & Martignon, L. (2002). Representation facilitates reasoning: what natural frequencies are and what they are not, *Cognition*, v. 84, issue 3, 343-352
- Kahneman, D. & Tversky, A. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky (Eds), *Judgment under uncertainty: Heuristics and biases* (pp. 153-160). Cambridge, England: Cambridge University Press.
- Lewis, C., & Keren, G. (1999). On the difficulties underlying Bayesian reasoning: comment on Gigerenzer and Hoffrage. *Psychological Review*, 106, 411–416.
- Macchi, L. (2000). Partitive Formulation of Information in Probabilistic Problems: Beyond Heuristics and Frequency Format Explanations. *Organizational Behavior and Human Decision Processes Vol. 82, No. 2*, 217–236
- Macchi, L. (1995). Pragmatic aspects of the base-rate fallacy. *Quarterly Journal of Experimental Psychology: A, Human Experimental Psychology Vol 48A(1)*, 188-207
- Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge University Press
- Snyder, R. E. (1966) Mammography: Contributions and limitations in the management of cancer of the breast. *Clinical Obstetrics and Gynecology*, 9, 207-220.
- Tenenbaum, J. B. & Griffiths, T. L. (2001) Structure learning in human causal induction. *Advances in Neural Information Processing Systems 13*
- Tversky, A. & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90, 293-315.
- Villejoubert, G. & Mandel, D. R. (2002). The inverse fallacy: An account of deviations from Bayes’s theorem and the additivity principle. *Memory and Cognition* 30 (2), 171-178