The Role of Causal Models in Reasoning Under Uncertainty

Tevye R. Krynski (tevye@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

Numerous studies of how people reason with statistical data suggest that human judgment often fails to approximate rational probabilistic (Bayesian) inference. We argue that a major source of error in these experiments may be misunderstanding causal structure. Most laboratory studies demonstrating probabilistic reasoning deficits fail to explain the causal relationships behind the statistics presented, or they suggest causal mechanisms that are not compatible with people's prior theories. We propose that human reasoning under uncertainty naturally operates over causal mental models, rather than pure statistical representations, and that statistical data typically support correct Bayesian inference only when they can be incorporated into a causal model consistent with people's theory of the relevant domain. We show that presenting people with questions that clearly explain an intuitively natural causal structure responsible for a set of statistical data significantly improves their performance. In particular, we describe two modifications to the standard medical diagnosis scenario that each eliminates the phenomenon of base-rate neglect, merely by clarifying the causal structure behind false-positive test results.

Introduction

Can people arrive at correct probability judgments after reading sufficient statistical data? Decades of experimental inquiry into intuitive statistical inference have documented the ways in which human judgment deviates from rational Bayesian norms. Examples include the phenomena of base-rate neglect (Kahneman & Tversky, 1982), the conjunction fallacy (Tversky & Kahneman, 1983), and deviations from the additivity principle (Villejoubert & Mandel, 2002). Yet in the real world, an environment that is saturated with useful statistical information and that continually poses challenges for reasoning under uncertainty, people function quite well, and far better than any artificial systems built on the norms of probability theory (Russell & Norvig, 2002).

One possible explanation for this discrepancy is that laboratory studies typically present participants with unnatural forms of information – single-event or epistemic probabilities instead of naturally sampled frequencies – and that human minds are only designed to operate on information in the latter, more natural format (Gigerenzer & Hoffrage, 1995). While we do not dispute the benefits of presenting people with statistical data in frequency formats, we doubt that the simple frequency-based algorithms of Gigerenzer and Hoffrage (1995) are responsible for most of our successful reasoning in everyday life. Real-world systems are too complex, and often sufficiently different from anything we have seen before, to support reasoning based on simply looking up frequencies in a table compiled from past experience.

Here we propose an alternative account of probabilistic reasoning errors in laboratory tasks, based on a different conception of how uncertain reasoning operates in the real world. We argue that human reasoning under uncertainty naturally operates over causal mental models, rather than purely statistical representations, and that statistical data typically support correct Bayesian inference only when they can be incorporated into a causal model consistent with people's theories of the domain. We will argue that misunderstanding causal structure is a major source of error in standard laboratory studies of probabilistic reasoning, and then describe two modifications to a standard task which are each capable of eliminating the typical "base-rate neglect" error by clarifying the causal structure of the problem.

Bayesian Inference and Base-rate Neglect

We focus on diagnostic reasoning problems: inferring the probability of a proposition H based on some observed data D. The normative Bayesian approach to diagnostic inference requires two kinds of probabilities: the prior, $P(H)$, representing our degree of belief that the hypothesis is true before making the observation, and the likelihoods, $P(D \mid H)$ and $P(D \mid \neg H)$, representing the probabilities that the data would have been observed if the hypothesis were true and if the hypothesis were false, respectively. Bayes' rule then prescribes an equation for computing the posterior, our degree of belief in the hypothesis given the data: $P(H \mid D) = P(H) \times P(D \mid H) / P(D)$, where $P(D)$ is computed as $P(H) \times P(D \mid H) + (1 - P(H)) \times P(D \mid \neg H)$.

Bayes' theorem does not prescribe how one should set the prior probability or the likelihoods, but most researchers have assumed that experimental participants should set them based on the statistics provided, specifically setting the prior equal to the base rate. The term "base rate" refers to a statistic summarizing how often H has been true in similar previous situations, independent of whether D was observed. The label "base-rate neglect" refers to errors in probabilistic reasoning that appear to be due to not setting $P(H)$ equal to the presented base rate, or to ignoring the influence of the $P(H)$ term in Bayes' rule.

Our primary example of base-rate neglect is a word problem adapted from Eddy (1982) and tested in an influential paper by Gigerenzer and Hoffrage (1995), and several follow-ups (Cosmides & Tooby, 1996; Lewis & Keren, 1999; Macchi, 2000). 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? ____ %

The probabilities here are called "single-event probabilities", or "epistemic probabilities"; they refer to degrees of belief about an individual case rather than the frequency of an outcome in a series of repeated trials. People are poor at solving this diagnosis problem, often giving an answer of 70%-90%, while Bayes' rule prescribes an answer of 7.8% (Gigerenzer & Hoffrage, 1995). Kahneman and Tversky (1982) used the term "base-rate neglect" to characterize errors in this range because in this and similar problems people seemed to be neglecting the base rate of 1% (the prior) in favor of the individuating information (the likelihoods), rather than combining the two via Bayes' rule to calculate the posterior. Other explanations for this phenomenon have been offered, such as the tendency to confuse a given conditional probability with its inverse (Villejoubert & Mandel, 2002). Macchi (1995) has catalogued incorrect answers to typical inference problems and found that most instances of "base-rate neglect" are best described as calculating $P(D \mid H)$, $1 - P(D \mid \neg H)$, or $P(D \mid H) - P(D \mid \neg H)$. Few participants actually carry out a Bayesian computation that neglects priors, which would produce answers equal to $P(D \mid H) / [P(D \mid H) + P(D \mid \neg H)]$.

Regardless of how one categorizes errors, one thing is clear: people do not possess a general-purpose probabilistic reasoning engine that takes as input single-event probabilities, sets priors and likelihoods equal to the corresponding statistics, and outputs correct posterior probabilities. But if people do not have such an ability, how are they generally able to navigate the world so well? Gigerenzer and Hoffrage (1995) propose that people do have the ability to make correct Bayesian computations, but typical laboratory problems present the statistical information in an unnatural format. They have shown that questions provided in a natural frequency format, rather than a probabilistic format, can dramatically reduce inference errors such as base-rate neglect. For instance, they tested the following "natural frequency" version of the mammography problem:

> 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 ___.

Gigerenzer and Hoffrage (1995) explain these results on evolutionary grounds, arguing that "as humans evolved, the 'natural' format was frequencies as opposed to probabilities or percentages." However, we know that people can use simple probabilities and percentages to reason correctly, and can often solve more complex probabilistic reasoning problems provided that the causal relevance of all factors is made clear (Kahneman & Tversky, 1980). The frequentist hypothesis does not explain success in these cases. Furthermore, the reduction of error in frequency formats could be due to the fact that Bayesian diagnosis problems phrased in terms of frequencies are just simpler to solve, involving only the addition of two whole numbers rather than the multiplication and division of six decimal numbers.

Gigerenzer and Hoffrage point to the simplicity of the frequency calculation as evidence that people only needed to evolve a simple inference system, but the success of the frequentist algorithm in simple word problems is not enough to justify the claim that people use this simple frequency computation for most real-world inferences. One never hears a mechanic say: "If you want to estimate the chances of your car breaking down on a long road trip, first think of the last 1000 cross-country road trips you took…." This approach only works when only one or two variables are relevant, and ample statistics are available. Real-world systems present complex patterns of correlation over many variables, and people typically do not have access to enough observations to warrant drawing conclusions based on a simple look-up table of frequencies of previous occurrences. Rather than appealing to a large collection of similar past experiences, we typically make judgments about the probability of a car breaking down, the chance of getting a certain job, or an acquaintance's intentions, by constructing and manipulating some kind of domain-specific causal mental model. This capacity to reason with causal mental models may also be responsible for our successes – and failures – on probabilistic reasoning tasks in the laboratory.

Bayesian Inference with Causal Models

Both the frequentist algorithms and standard Bayesian inference are domain-general and purely statistical approaches to uncertain reasoning. We propose that rather than possessing a domain-general engine taking statistical data as input and producing probabilities of hypotheses as output, people naturally evaluate and interpret statistical information within the framework of a domain-specific probabilistic causal model, derived from a theory of how particular kinds of causes produce particular kinds of effects in that domain. An individual's probabilistic causal model encompasses knowledge of which causes produce which effects (the structure), how likely certain causes are to occur (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 causal reasoning engine, which takes as input (1) a probabilistic causal model and (2) observations or statistical data, and is capable of producing probabilities of hypotheses as output. The causal reasoning engine can be formally modeled using the tools of Bayesian networks (Pearl, 2000), but for the purposes of this short paper, we limit our discussion to informal graphical representations of probabilistic causal models.

Graphical models have figured in many recent accounts of human categorization (Rehder, 2001; Waldmann et. al, 1995) and causal structure learning (Ahn & Dennis, 2000; Gopnik et al, in press; Steyvers, Tenenbaum et al., in press; Tenenbaum & Griffiths, 2001), but have not to date made a large impact on the study of reasoning under uncertainty

more generally. Yet the connections between real-world causality and uncertainty run deep – so deep that we doubt there can be a complete theory of reasoning under uncertainty that does not include, and perhaps center around, causality. Pearl (2000) argues that much of the uncertainty of inference in an otherwise deterministic world is due to multiple causal influences that can produce the same effect. For instance, coffee is occasionally bitter, but this is not due to a stochastic mechanism that unpredictably makes coffee bitter; rather it is due to one of several hidden causal influences: over-roasting or burning the coffee. If one wishes to know the probability that a given cup of coffee will be bitter, the first step should be to identify the potential causes of bitterness and then to investigate them, (e.g., how long has the pitcher been on the burner, etc.), rather than to start with a statistical analysis, e.g., estimating the proportion of bitter cups of coffee you've had in your lifetime. As all statistical correlations are ultimately a result of (perhaps very indirect) differential causal influences, we expect reasoning under uncertainty to be sensitive to the causal structures that create uncertainty in the first place.

The connection between causality and uncertain reasoning was one of many directions pioneered by Tversky and Kahneman. Tversky and Kahneman (1980) found that providing "causally relevant" base rates improved probabilistic inference, but they did not explain why, or even define what they meant by "causally relevant". Their most explicit proposal was that "base-rate information which is not incorporated into a causal schema, either because it is not interpretable as an indication of propensity or because it conflicts with an established schema, is given little or no weight." Tversky and Kahneman treated causal schemas as potential sources of error in statistical reasoning, whereas we take them as necessary substrates for probabilistic inference to succeed in complex, everyday scenarios. The effects of causal schemas are not indicators of how some "pure" statistical reasoning engine may go wrong, but the sign that people are not doing "pure" statistical reasoning at all; they are doing intrinsically causal reasoning, by computing probabilities over causal mental models.

Our goal is to go beyond the notion of "causally relevant" base rates by examining more precisely how causal mental models provide the substrate for reasoning under uncertainty, and how those models are constructed. We view causal models as transient mental representations constructed on the fly to solve specific problems, based on both given information and the constraints imposed by people's domain theories. For instance, one's theory of electricity should not allow one to construct a causal model in which taking the batteries out of a device causes it to start working. Our evidence suggests that any piece of given information – base rates, likelihoods, or qualitative statements – will only be used if people can incorporate it into a causal model compatible with their domain theory.

More specifically, we will 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. Based on the information provided in the problem, people may assume that false positives are caused by noise or random error. Since doctors presumably trust the test, people might further assume that the level of noise is low, and this assumption is incompatible with the statistics provided. In fact, the statistics provided are not compatible with the actual causal structure of standard mammogram screenings.

Gigerenzer and Hoffrage adapted the probabilistic version of the breast cancer problem from Eddy (1982), which describes the true statistical nature of mammograms. We found several important discrepancies between the true statistics in Eddy (1982) and those presented to participants by Gigerenzer and Hoffrage (1995), which could be at least partly responsible for their participants' poor performance.

1. In Eddy's paper, the likelihoods of 80% and 9.6% are 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 chose to apply the likelihood of 9.6% to all women without cancer, rather than just those with benign lesions. Participants thus had no indication that benign lesions are actually the cause of the false positives.
2. The structure of the problem is misleading, by simply 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.

As Gigerenzer and Hoffrage have described it, the mammogram appears to be an extremely error-prone test: 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 a test, with a noise rate 10 times higher than the base rate of cancer (1%)? We believe a principal reason people perform so poorly is that they have difficulty understanding how such a high false-alarm rate could result purely from noise (the only cause of a false alarm they are aware of) given that doctors trust this test enough to declare the result "positive".

Probabilistic Causal Models

A basic causal model for this scenario is depicted in Figure 1A, in which a positive mammogram can result from one of two independent and stochastic causes: the patient having cancer or the test having noise. Formally, this model can be represented as a Bayes net with a noisy-or parameterization (Cheng, 1997; Pearl, 2000). If there were only one potential cause, the probability that the effect occurs is just the base

rate of the cause times the *causal power* of that cause (a conditional probability, between 0 and 1; Cheng, 1997); with multiple potential causes, the probability that the effect occurs is equal to the probability that one or more of its causes occurs and succeeds in causing the effect (treating both the occurrence of causes and their causal powers as independent).

Suppose a participant believes a positive mammogram to be tantamount to a doctor's diagnosis of breast cancer. This is not implausible: doctors as a rule avoid scaring patients unnecessarily, and it is common for them to say, "You have some indications consistent with disease X, but it's probably nothing". If instead the doctor says, "You've tested positive for breast cancer," there should be a good chance that you actually have cancer. In this case, people may assume that the base rate of noise would not be higher than the base rate of cancer. Otherwise, the doctor would say, "It could be cancer, but there's a good chance it was just noise". This assumption, however, is inconsistent with the high 9.6% false-positive rate and the causal model described above; at most, the false alarm rate could equal the base rate of cancer (1%). People reasoning with this model could become confused at this point and just look for some way to combine the given numbers to obtain a reasonable estimate.

As discussed above, the model in Figure 1A does not reflect the true causal structure of the test. Figure 1B shows a more realistic model, in which the source of the false positives is an alternative tissue anomaly: dense benign lesions. Now, the 9.6% false-positive rate can be naturally interpreted as the approximate base rate (for women with a breast mass) of having a benign lesion dense enough to cause a positive mammogram. This interpretation is perfectly consistent with people's background knowledge that tissue anomalies (e.g., pimples, moles, birthmarks, or bumps) are often harmless.

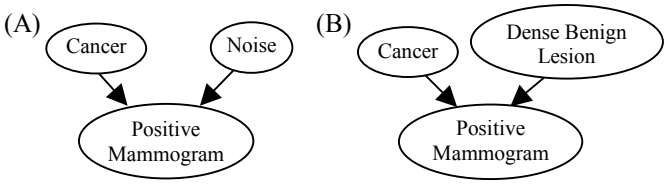

Figure 1: (A) basic causal model of mammogram, with noise. (B) more accurate model with specific alternative cause.

Experiment 1

Our first experiment directly tested the idea that people might tacitly assume a positive result to be tantamount to a doctor's diagnosis of cancer. We hypothesized that people would better understand the given statistics if most women without cancer who did not test "negative" received an "uncertain" result rather than a "positive" one. Since a doctor's report of "uncertain" implies that she believes the test outcome could well be the result of random noise, participants could naturally incorporate the 9.6% "uncertain" rate in healthy women as the base rate of the noise variable in Figure 1A.

Method

Participants. 73 airplane passengers were recruited while waiting for their flights to begin boarding. Their only compensation was temporary alleviation of boredom.

Design. Participants were given paper-and-pen versions of Gigerenzer's breast cancer question, with the modification that the test has three possible results: "positive", "uncertain", and "negative" (inspired by Eddy, 1982). Participants received one of two versions: in one, a woman gets a "positive" result; in the other she gets an "uncertain" result. The numbers were exactly the same in both versions, except that the conditional probabilities for "positive" and "uncertain" were switched, so the same calculations were required in both versions. The questions follow:

"Positive" Question

Women at age 40 are often encouraged by their doctor to participate in a routine mammography screening for breast cancer. The mammogram has 3 possible results:

Positive: the patient has breast cancer. This results when tumors are found that are definitely cancerous.

Uncertain: the patient may have breast cancer. This result occurs when tissue exists that may be normal breast tissue, benign tumor, or cancerous tumor. More testing is needed to determine whether the patient has breast cancer.

Negative: the patient does not have breast cancer.

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

[first 14 lines identical to "positive" question]

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 and Discussion

A one-way ANOVA of the raw responses revealed a significant difference between the two versions ($F(1,71)= 21.59$, $MSE=897.36$, $p<.0001$). We classified as base-rate neglect any answer greater than or equal to 70%. Since the exact correct answer of 7.8% is difficult to calculate, we classified as "close" any answer between 5% and 12% inclusive. We also classified answers of 1% or 2% as base rate overuse. The result was a significant difference between the two versions ($\chi^2(3) = 16.15$, $p < .005$).

Table 1: "Positive" versus "Uncertain" Questions

Mammogram	Base-rate Neglect	Close Answer	Base-rate Overuse	Other
Positive	14	9	6	6
Uncertain	1	15	14	8

These results are consistent with our hypothesis that "base-rate neglect" may arise in the basic question because people take the "positive" label to mean that the doctor trusts the test, thus limiting the base rate of noise to a level inconsistent with the high false-alarm rate. An alternative interpretation is that participants are answering simply based on the meaning of the words "uncertain" and "positive", rather than reasoning about likely levels of noise. To test this, we gave 36 new participants a third "control" question in which we relabeled the "positive" result "certain" and the "uncertain" result "positive", so that "positive" now means the patient may have cancer, and more testing is needed:

"Control" Question

The mammogram has 3 possible results:

Certain: the patient has breast cancer. This results when tumors are found that are definitely cancerous.

Positive: the patient may have breast cancer. This result occurs when tissue exists that may be normal breast tissue, benign tumor, or cancerous tumor. More testing is needed to determine whether the patient has breast cancer.

[the rest of the question is identical to the "positive" question]

The incidence of base-rate neglect for this "control" question was significantly higher than in the "uncertain" question (6/36 versus 1/38, $\chi^2(1) = 4.25$, $p < .05$), but much lower than in the "positive" question (6/36 versus 14/35, $\chi^2(1) = 4.78$, $p < .05$). The only difference in the latter case was defining "positive" as "may have cancer" rather than "has cancer". This result suggests that "base-rate neglect" is due to a mismatch between the information given (a false-positive rate much higher than the cancer rate) and people's domain knowledge (the only cause of false positives they are aware of is noise, and a trusted test implies a noise rate lower than the disease rate).

Experiment 2

While Experiment 1 focused on constraints imposed by domain knowledge, Experiment 2 directly tests the role of causal reasoning. Specifically, we investigated whether people would more easily integrate the high false-positive rate into their causal models if they knew what causes false positives: dense benign cysts. They could then use the causal model of Figure 1B, assimilating the high false-positive rate as the base rate of an alternative kind of tissue anomaly, which would not be inconsistent with their domain knowledge.

Method

Participants. 155 people were recruited at the airport or the MIT campus. MIT students were compensated with candy; airplane passengers were compensated as in Experiment 1.

Design. We posed two paper-and-pen questions, one with only statistical information about false positives and one with information about an alternative cause for a positive result. Crucially, both versions required the exact same Bayesian formula to calculate the answer. To minimize arithmetic errors, participants were allowed to answer with either ratios or percentages. We also varied the base rate and false-positive likelihoods (1% and 5% respectively vs. 2% and 6%), and the cover story (breast cancer and harmless cyst vs. colon cancer and harmless polyp), for a total of 8 different questions. Sample questions were as follows (for variants, see http://web.mit.edu/tevya/www/CogSci20003):

"Statistical" 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 specify the true positive rate, but only that "most women with breast cancer will receive a positive mammogram." We made this change to encourage participants to provide answers based on their intuition rather than memorized mathematical formulas.

Results and Discussion

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 raw responses showed no significant interactions, with a significant difference between "Statistical" and "Causal" questions (F=8.33, p<.005), and no significant effect of cover story (F=0.43, p=.51) or prior and false-positive likelihood values (F=.0052, p=.94), all with df=(1,125), MSE=836. We classified as base-rate neglect any answer greater than or equal to 70%. We classified as correct any answer equal to between 80% and 100% of the correct ratio or percentage. (This range accommodates the fact that most, but not all, women with cancer receive positive results.) The causal version significantly reduced base-rate neglect and improved correct responding as compared to the statistical version ($\chi^2(2) = 12.83$, $p < .0005$) (see Table 2).

Table 2: "Statistical" versus "Causal" Questions

Problem Type	Base-rate Neglect	Correct (or close)	Base-rate Overuse	Other
Statistical	19	24	17	16
Causal	3	40	16	20

General Discussion

In two experiments, we gave people a natural way to make sense of the high false-positive rate in terms of their causal mental models of the mammogram scenario, and thereby essentially eliminated the phenomenon of base-rate neglect. A total of 4 out of 117 participants exhibited base-rate neglect on our new questions, compared to 33 out of 111 people on questions paralleling the original version, despite the required calculations being identical. Likewise, the incidence of correct or near-correct responses increased from 33 out of 111 participants to 55 out of 117. Experiment 1 showed that diagnostic reasoning could be improved by removing the inconsistency between an apparently high noise rate and an apparently trusted test. Experiment 2 showed that reasoning could be improved by introducing a compelling non-noise alternative cause for the frequent false positives. We interpreted these findings as evidence that human probabilistic reasoning operates over causal mental models rather than purely statistical databases. We also argued that this central role for causality in reasoning under uncertainty should be considered rational and normative, contrary to standard assumptions in the Heuristics and Biases (Kahneman & Tversky, 1982) or Natural Frequency research programs (Gigerenzer & Hoffrage, 1995).

From the standpoint of probabilistic causal models, the real problem behind "base-rate neglect" errors comes not from having a low base rate for the cause in question, but from having a high false-positive rate. Assuming independent, probabilistically sufficient causes, as in the noisy-or model, and assuming that each cause is relatively rare, suggests a natural interpretation for the true positive rate $P(D \mid H)$ in terms of the approximate causal strength of H. But the false-positive rate $P(D \mid \neg H)$, while just as important as the true positive rate in purely probabilistic reasoning, has no such natural causal interpretation; an effect cannot result from the absence of a cause. In causal reasoning, we must come up with one or more alternative causes to account for false positives. Whether that can be done coherently depends on the match between the statistics given in the problem and our intuitive domain theories, which determine what alternative causes are likely to be considered and constrain their base rates and causal strengths. Telling people about an alternate cause whose base rate could plausibly be high enough to account for the given false alarm rate, such as the "dense benign cysts" in the breast cancer scenario, could thus make a huge contribution to improving uncertain reasoning.

Despite the advantages of probabilistic causal reasoning over purely statistical reasoning, the successes of real-world inference cannot be explained just by appealing to causal models. In order to construct a causal model for a given scenario, people must recruit domain-specific theories that specify which kinds of causes are likely to produce which kinds of effects. But what does that theoretical knowledge consist of, and how is it used to constrain causal model construction? Understanding how causal models are constructed through the interaction of domain theories (top-down constraints) and statistical data (bottom-up constraints) is a largely open question, and the answer should play a critical role in explaining how people reason so successfully and efficiently in an uncertain world.

References

Ahn, W., & Dennis, M. (2000). Induction of causal chains. *Proceedings of the Twenty-second Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review. Vol 104(2)*, 367-405.

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*, 1-73.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge..

Gigerenzer, G. & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*, 684-704

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. *Psych. Review*.

Kahneman, D. & Tversky, A. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky (Eds), *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge.

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*, *82*, 217–236.

Macchi, L. (1995). Pragmatic aspects of the base-rate fallacy. *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *48A(1)*, 188-207.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Russell, S., & Norvig, P., (1995). *Artificial Intelligence: a Modern Approach*. Prentice Hall.

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*. MIT Press.

Tversky, A. & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*, 293-315.

Tversky, A. & Kahneman, D. (1980). Causal schemas in judgments under uncertainty. In M. Fishbein (Ed.), *Progress in Social Psychology*. Mahwah, NJ: Erlbaum.

Waldmann, M. R., Holyoak, K. J., & Fratianne, A. (1995). Causal models and the acquisition of category structure. *Journal of Experimental Psychology: General, 124*, 181-206.

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

Acknowledgements We thank Liz Baraff for helping with experiments and Tom Griffiths for statistical assistance.