The Impact of Information Representation on Bayesian Reasoning

Ulrich Hoffrage and **Gerd Gigerenzer**
Center for Adaptive Behavior and Cognition
Max Planck Institute for Psychological Research
Leopoldstrasse 24, 80802 Munich, Germany.
{hoffrage,gigerenzer}@mpipf-muenchen.mpg.de

Abstract

Previous research on Bayesian inference, reporting poor performance by students and experts alike, has often led to the conclusion that the mind lacks the appropriate cognitive algorithm. We argue that this conclusion is unjustified because it does not take into account the information format in which this cognitive algorithm is designed to operate. We demonstrate that a Bayesian algorithm is computationally simpler when the information is represented in a frequency rather than a probability format that has been used in previous research. A frequency format corresponds to the way information is acquired in natural sampling--sequentially and without constraints on which observations will be included in the sample. Based on the assumption that performance will reflect computational complexity, we predict that a frequency format yields more Bayesian solutions than a probability format. We tested this prediction in a study conducted with 48 physicians. Using outcome and process analysis, we categorized their individual solutions as Bayesian or non-Bayesian. When information was presented in the frequency format, 46% of their inferences were obtained by a Bayesian algorithm, as compared to only 10% when the problems were presented in the probability format. We discuss the impact of our results on teaching statistical reasoning.

Is the mind, by design, predisposed against performing Bayesian inference? The classical probabilists of the Enlightenment, including Condorcet, Poisson, and Laplace, who equated probability theory with the common sense of educated people, would have said the answer is no. And when Ward Edwards and his colleagues (Edwards, 1968) started to test experimentally whether human inference follows Bayes' theorem, they gave the same answer: although "conservative," inferences were usually proportional to those calculated from Bayes' theorem. Kahneman and Tversky (1972, p. 450), however, arrived at the opposite conclusion: "In his evaluation of evidence, man is apparently not a conservative Bayesian: he is not a Bayesian at all." In the 1970s and '80s, proponents of their "heuristics-and-biases" program amassed an apparently damning body of evidence that people systematically neglect base rates in Bayesian inference problems. This could be shown not only with students, but also with experts in their fields, for instance, with physicians (Casscells, Schoenberger, & Grayboys, 1978; Eddy, 1982).

Thus, there are two contradictory claims as to whether people naturally reason according to Bayesian inference. In this paper we argue that both views are based on an incomplete analysis: They focus on cognitive processes, Bayesian or otherwise, without making the connection between what we will call a cognitive algorithm and an information format. We (a) provide a theoretical framework (based on Gigerenzer and Hoffrage, 1995) that specifies why a frequency format should improve Bayesian reasoning and (b) present a study that tests this hypothesis.

Algorithms Are Designed for Information Formats

Our argument centers on the intimate relationship between a cognitive algorithm and an information format. This point was made in a more general form by the physicist Richard Feynman. In his classic The Character of Physical Law (1967), Feynman places great emphasis on the importance of deriving different formulations for the same physical law, even if they are mathematically equivalent (e.g., Newton's law, the local field method, and the minimum principle). Different representations of a physical law, Feynman reminds us, can evoke varied mental pictures and thus assist in making new discoveries: "Psychologically they are different because they are completely unequivalent when you are trying to guess new laws" (p. 53). Likewise, Stephen Palmer (1978) points out in his analysis of different modes of representation "that no form of representational equivalence guarantees that performance characteristics will be the same for two representations embedded in process models" (p. 272).

Consider numerical information as one example of an external representation. Numbers can be represented in Roman, Arabic, and binary systems, among others. These representations can be mapped one-to-one onto each other and are in this sense mathematically equivalent. But the form of representation can make a difference for an algorithm that does, say, multiplication. The algorithms of our pocket calculators, for instance, are tuned to Arabic numbers as input data and would fail badly if one entered binary numbers.

Our general argument is that mathematically equivalent representations of information entail algorithms that are not necessarily computationally equivalent. This point has an important corollary for research on inductive reasoning. Suppose we are interested in figuring out what algorithm a system uses. We will not detect the algorithm if the representation of information we provide the system does not match the representation with which the algorithm works. For instance, assume that in an effort to find out whether a system, such as a pocket calculator, has an algorithm for multiplication, we feed that system binary numerals. The observation that the system produces mostly garbage does not entail the conclusion that it lacks an algorithm for multiplication. We will now apply this argument to Bayesian inference.

Probability Format

In this paper we focus on an elementary form of Bayesian inference that has been the subject of almost all experimental studies on Bayesian inference in the last 25 years. The following "mammography problem" (adapted from Eddy, 1982) is one example:

> Mammography problem (probability format)
> The probability of breast cancer is 1% for a woman at age forty who participates in routine screening. If a woman has breast cancer, the probability is 80% that she will have a positive mammography. If a woman does not have breast cancer, the probability is 10% that she will also have 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? _____%

There are two mutually exclusive and exhaustive hypotheses (H: breast cancer and –H: no breast cancer) and one observation (D: positive test). All information (base rate, hit rate, and false alarm rate) is represented in terms of single-event probabilities attached to a single person. (Here, they are expressed as percentages; alternatively, they can be presented as numbers between zero and one.) The task is to estimate a single-event probability. The algorithm needed to calculate the Bayesian posterior probability p(cancer|positive) from this format can be seen in Figure 1 (left side), where the information is already inserted into Bayes' rule. The result is .075.

We know from several studies that physicians, college students (Eddy, 1982), and staff at Harvard Medical School (Casscells, Schoenberger, & Grayboys, 1978) all have equally great difficulties with this and similar medical disease problems. For instance, Eddy (1982) reported that 95 out of 100 physicians estimated the posterior probability to be between 70% and 80%, rather than 7.5%.

In the last few decades, this probability format has become a common way to communicate information, found everywhere from medical and statistical textbooks to psychological experiments. Not surprisingly, the experimenters who have amassed the evidence that humans fail to meet the norms of Bayesian inference have usually given their subjects information in the probability format (or its variant, in which one or more of the three percentages are relative frequencies). But it is only one of many mathematically equivalent ways of representing information. It is, moreover, a recently invented notation: Percentages became common notation only during the 19th century. How did organisms acquire information before that time?

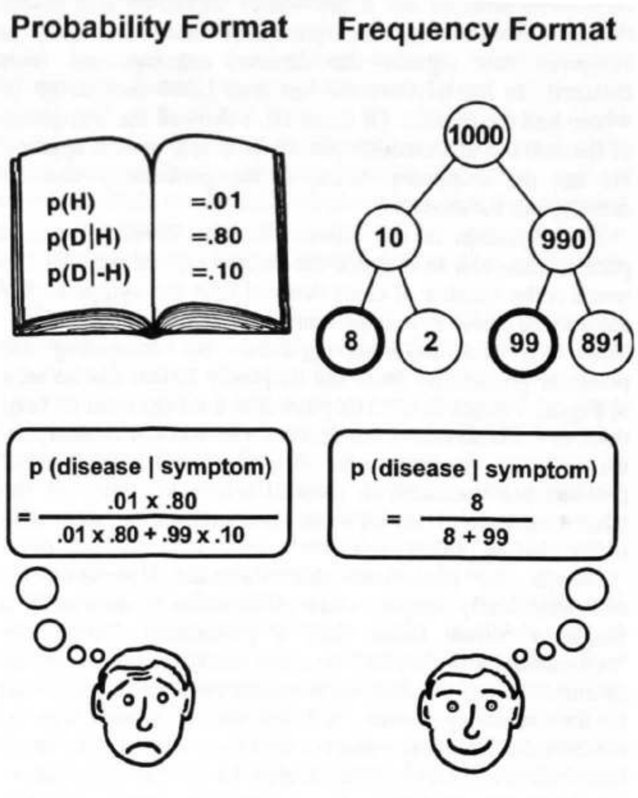

Figure 1: Bayesian inference and information representation (probability format and frequency format with frequencies as obtained by natural sampling).

Natural Sampling of Frequencies

We assume that as humans evolved, the "natural" format was frequency as actually experienced in a series of events, rather than probability or percentage. From animals to neural networks, systems seem to learn about contingencies through sequential encoding and updating of event frequencies. This sequential acquisition of information by updating event frequencies without artificially fixing the marginal frequencies (e.g., of disease and no-disease cases) is what we refer to as natural sampling (Kleiter, 1994). Brunswik's "representative design" is a special case of natural sampling (Brunswik, 1955). In contrast, in experimental research the marginal frequencies are typically fixed a priori. For instance, an experimenter may want to investigate 100 people with disease and a control group of 100 people without disease. This kind of sampling with fixed marginal frequencies is not what we refer to as natural sampling.

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. Imagine an old, experienced physician in an illiterate society. She has no books or statistical surveys and therefore must rely solely on her experience. Her people have been afflicted by a previously unknown and severe disease. Fortunately, the physician has discovered a symptom that signals the disease, although not with certainty. In her lifetime she has seen 1,000 people, 10 of whom had the disease. Of those 10, 8 showed the symptom; of the 990 not afflicted, 95 did. Now a new patient appears. He has the symptom. What is the probability that he actually has the disease?

The physician in the illiterate society does not need a pocket calculator to estimate the Bayesian posterior. All she needs is the number of cases that had both the symptom and the disease (here: 8) and the number of symptom cases (here: 8 + 95). The Bayesian algorithm for computing the posterior probability from the frequency format can be seen in Figure 1 (right side). The physician does not need to keep track of the base rate of the disease. Her modern counterpart, the medical student who struggles with single-event probabilities presented in medical textbooks, may on the other hand have to rely on a calculator and end up with little understanding of the result.

So far, we have seen that Bayesian algorithms are computationally simpler when information is encoded in a frequency format rather than a probability format. By "computationally simpler" we mean that (a) fewer operations (multiplication, addition, or division) need to be performed in the frequency format, and (b) the operations can be performed on natural numbers (absolute frequencies) rather than fractions (such as percentages). From this observation, we derive the prediction that a frequency format elicits a substantially higher proportion of Bayesian algorithms than a probability format. Henceforth, when we use the term "frequency format," we always refer to frequencies as defined by the natural sampling tree in Figure 1.

Study: Frequency Formats Improve Bayesian Reasoning

In a study previously conducted with 60 subjects from the University of Salzburg, Austria (see Gigerenzer & Hoffrage, 1995, Study 1), we demonstrated that the frequency format elicited a substantially higher proportion of Bayesian algorithms than the probability format. In 15 different inferential problems, including the mammography problem, Bayesian reasoning went up from 16% in the probability format to 46% and 50% in two versions of the frequency format. No instruction or feedback was given; the information format by itself improved Bayesian reasoning. Similar results were obtained by Christensen-Szalanski and Beach (1982) and Cosmides and Tooby (1996). Now, remember that both Casscells et al. (1978) and Eddy (1982) reported poor performances from the physicians they investigated. Because Bayesian reasoning is of great importance in medicine, the goal of the current study was to see whether not only students but also physicians could gain from a frequentistic representation of the information. One might suspect that this method only works with students who lack experience in diagnostic inference, but not with physicians who make diagnostic inferences every day. On the other hand, medical textbooks typically present information about base rates, hit rates, and false alarm rates in a probability format (as in Figure 1, left side). Just as a pocket calculator is unable to process binary numerals adequately, physicians may be unable to process statistical information if it is presented in a format for which their minds were not designed.

Method

Participants. We investigated 48 Munich physicians, 18 from university hospitals, 16 from private or public hospitals, and 14 from private practice. Mean age was 42 years and mean time of professional service was 14 years with a range of one month to 30 years (our sample included beginners as well as directors of clinics). They were studied individually.

Materials. We used four medical problems, including the mammography problem adapted from Eddy (1982). The other three problems concerned (1) colon cancer and positive haemoccult blood test, (2) Bechterew's disease and HL-Antigen B27, and (3) Phenylketonuria and positive Guthrie-test as disease and symptom, respectively. We consulted experts and the literature to determine the best statistical information available for the base rates, hit rates, and false alarm rates.

Design and Procedure. For each problem we constructed two versions: one in the probability format and one in the frequency format. Participants received a booklet containing all four problems, two of them in probability format and two in frequency format. Assignment of problems to formats, as well as order of formats and problems were completely counterbalanced.

The physicians worked on the booklet at their own pace (on average 7 minutes per problem). Each problem was on one sheet, followed by a separate sheet where the physicians were asked to make notes, calculations, or drawings. After filling out the booklet they were interviewed about their mental processes.

Results

We classified an inferential process as a Bayesian algorithm only if (a) the estimated probability or frequency was the same as the value calculated from applying Bayes' theorem to the information given (outcome criterion), and the notes the physicians made while solving the problems and/or the follow-up interviews suggested that the answer was not just a guess but a Bayesian computation as defined by the equations in Figure 1 (process criterion), or if (b) the solution was obtained by a shortcut algorithm that still provided the correct answer plus or minus 5 percentage points.

The results confirmed our prediction: Across all 96 individual problem solutions for the probability format (48 physicians times two problems), 10% were correct, whereas for the frequency format, 46% were correct ($t_{df=190}=5.5$, $p<0.001$). Figure 2 shows the absolute frequencies of Bayesian solutions for the four problems.

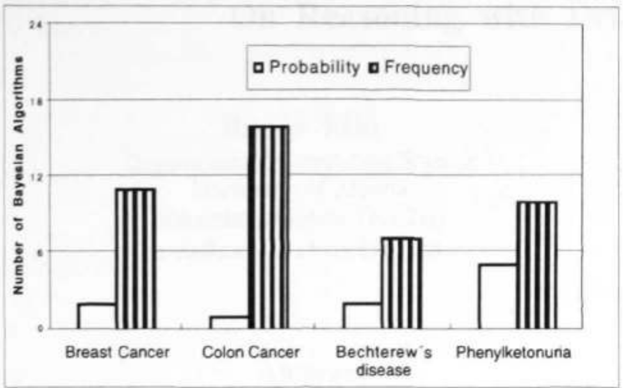

Figure 2: Number of Bayesian algorithms in the four problems. (Maximum number possible: 24.)

This difference in performance is reflected in the remarks the physicians made while working on the problems. For instance, when working on probability-format problems, several made complaints such as "I simply can't do that. Mathematics is not my forte," or "There is a formula, but at the moment I can't derive it." However, with a frequency format, some typical remarks were, "Now it's different. It's quite easy to imagine. There is a frequency; that's more visual," or "Oh, how nice--this is just like the word problems we did in elementary school. A first grader could do this. Wow, if someone can't solve this...!" Like the Bayesian algorithms, the non-Bayesian algorithms were also format-specific: In 18 (5) out of the 96 probability (frequency) versions, our physicians gave the hit rate, p(D|H), as the posterior. For the algorithm that we termed likelihood subtraction, p(D|H) p(D|-H), the corresponding numbers were 20 (5) out of 96. Two of the algorithms that were dominant in the frequency format were base rate only, p(H), which was applied in 1 (15) out of 96 cases in the probability (frequency) format, and percentage positive, p(D), where frequency of use was 0 (9) out of 96, respectively. (Less frequent algorithms are not reported here). For 28 (12) out of the 96 problem solutions in the probability (frequency) version we were unable to identify any algorithm at all.

The physicians spent about 25% more time on the probability problems, which reflects that they found these more difficult to solve. Many of them reacted -- cognitively, emotionally, and physiologically -- differently to probability and frequency formats. They were more often nervous when information was presented in probabilities, and they were less skeptical of the relevance of statistical information to medical diagnosis when the information was in frequencies. Bayesian responses were age correlated: The older half of the physicians (more than 40 years old) contributed only 37% of the Bayesian solutions, the younger half 63%.

Discussion

We return to our initial question: Is the mind, by design, predisposed against performing Bayesian inference? The conclusion of 25 years of heuristics-and-biases research would suggest as much. This previous research, however, has consistently neglected the insight that mathematically equivalent information formats need not be psychologically equivalent. An evolutionary point of view suggests that the mind is tuned to a frequency format, which is the information format humans encountered long before the advent of probability theory. We have shown that mathematically equivalent representations of information can entail computationally different Bayesian algorithms and we reported a study conducted with physicians that demonstrated how performance can be improved by presenting the information in the frequency rather than the probability format.

This striking result can be useful for teaching statistical reasoning--a field that is still neglected, not only in high school mathematics education but often in research as well. Up until now, only a few studies have attempted to teach Bayesian inference, mainly by outcome feedback, and with little or no success. The present framework suggests an effective way to teach Bayesian inference and statistical reasoning in general: Instead of teaching rules and how to insert probabilities into them, it seems to be more promising to teach representations and how to translate probabilities into frequency representations. Sedlmeier and Gigerenzer (1996) implemented both methods in a computerized tutorial system. And indeed they could show that teaching representations yielded performances more than twice as good as those obtained by rule training. Moreover, the advantage remained stable 5 weeks after training, whereas the effect of the rule-learning program had shown the usual rapid decay.

However, besides teaching statistical reasoning, there is a much more direct impact of our results. Physicians are often reported to become uneasy or even angry when asked for statistical information (Eddy, 1988), and to believe that their patients do not understand, or do not want to understand, the uncertainties inherent in diagnosis and therapy (Katz, 1988). We imagine that a frequency format might help improve the communication between patients and physicians (Bursztajn et al., 1981) and provide a tool for helping the patient to become a more apt decision maker.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (Ho 1847/1-1). We thank Maria Zumbeel for collecting the data and Anita Todd for editing the manuscript.

References

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. Psychological Review, 62, 193-217.

Bursztajn, H., Feinbloom, R. I., Hamm, R. H., & Brodsky, A. (1981). Medical choices, medical chances. How patients, families, and physicians can cope with uncertainty. New York: Delta.

Casscells, W., Schoenberger, A., & Grayboys, T. (1978). Interpretation by physicians of clinical laboratory results. New England Journal of Medicine, 299, 999-1000.

Christensen-Szalanski, J. J. J., & Beach, L. R. (1982). Experience and the base-rate fallacy. Organizational Behavior and Human Performance, 29, 270-278.

Cosmides, L., & 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 (pp. 249-267). Cambridge: Cambridge University Press.

Eddy, D. M. (1988). Variations in physician practice: the role of uncertainty. In J. Dowie & A. S. Elstein (Eds.), Professional judgment. A reader in clinical decision making (pp. 45-59). Cambridge: Cambridge University Press.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), Formal representation of human judgment (pp. 17-52). New York: Wiley.

Feynman, R. (1967). The character of physical law. Cambridge, MA: MIT Press.

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

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. Cognitive Psychology, 3, 430-454.

Katz, J. (1988). Why doctors don't disclose uncertainty. In J. Dowie & A. S. Elstein (Eds.), Professional judgment. A reader in clinical decision making (pp. 544-565). Cambridge: Cambridge University Press.

Kleiter, G. D. (1994). Natural sampling: Rationality without base rates. In G. H. Fischer & D. Laming (Eds.), Contributions to mathematical psychology, psychometrics, and methodology (pp. 375-388). New York: Springer.

Palmer, S. E. (1978). Fundamental aspects of cognitive representation. In E. Rosch & B. B. Lloyd (Eds.), Cognition and Categorization (pp. 262-303). New York: Lawrence Erlbaum.

Sedlmeier, P., & Gigerenzer, G. (1996). Teaching Bayesian reasoning in less than two hours. Manuscript submitted for publication.