The psychology of Bayesian reasoning

David R. Mandel[*]
Socio-Cognitive Systems Section, Defence Research and Development Canada and Department of Psychology, York University, Toronto, Ontario, Canada

[*] **Correspondence:** David R. Mandel, Socio-Cognitive Systems Section, Defence Research and Development Canada, 1133 Sheppard Avenue West, Toronto, Ontario, M3K 2C9, Canada
david.mandel@drdc-rddc.gc.ca

Keywords: Bayesian reasoning, belief revision, subjective probability, human judgment, psychological methods.

Most psychological research on Bayesian reasoning since the 1970s has used a type of problem that tests a certain kind of statistical reasoning performance. The subject is given statistical facts within a hypothetical scenario. Those facts include a base-rate statistic and one or two diagnostic probabilities. The subject is meant to use that information to arrive at a "posterior" probability estimate. For instance, in one well-known problem (Eddy, 1982) the subject encounters the following:

> 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 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 information in such problems can be mapped onto common expressions that use H as the focal hypothesis, $\neg H$ as the mutually-exclusive hypothesis, and D as datum: $P(H)$, the prior (often equated with the base-rate) probability of the hypothesis; $P(D|H)$, the true-positive rate; and $P(D|\neg H)$, the false-positive rate. In the mammography problem, $P(H) = .01$, $P(D|H) = .80$, and $P(D|\neg H) = .096$. Furthermore, $P(\neg H) = 1 - P(H) = .99$. The estimate queried is $P(H|D)$.

Bayes' theorem states:

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

Thus it yields a posterior probability of .078 in the mammography problem. Yet even the majority of physicians who were queried by Eddy (1982) gave estimates roughly one order of magnitude higher (i.e., .70-.80).

Well-established findings such as these have supported the view that expert and naïve subjects alike are non-Bayesian (Kahneman and Tversky, 1972). A common explanation is that people neglect base-rate information, which is not tracked by the intuitive heuristics they use to reach an estimate (Tversky and Kahneman, 1972, 1973). For instance, if base rates were neglected in the mammography problem,

$$P(H \mid D) = \frac{.80}{.80 + .096} \approx .89.$$

This estimate is closer to the modal estimate but is still off by about ten percentage points. Another explanation is that people commit the *inverse fallacy*, confusing $P(H|D)$, which they are asked to estimate, with $P(D|H)$, which is provided (Koehler, 1996). In the mammography problem, this explanation fits the data well because $P(D|H) = .80$. The inverse fallacy can also explain patterns of deviation from Bayes' theorem in tasks that hold constant base rates for alternative hypotheses (Villejoubert and Mandel, 2002).

It is also known that steps can be taken to increase agreement with Bayes' theorem. Since Bayes' theorem can be simplified as

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

task reformulations that directly provide these values or make them easily computable increase the proportion of Bayesian responses (e.g., Ayal and Beyth-Marom, 2014; Gigerenzer and Hoffrage, 1995; Hoffrage et al., 2002). Such formulations of evidence reduce computational steps and may also effectively trigger awareness of the correct solution, much as eliciting logically-related probability estimates (e.g., of binary complements) in close proximity rather than far apart improves adherence to the additivity property (Karvetski et al., 2013; Mandel, 2005). Natural frequency representations, which reveal nested-set relations among a reference class or representative sample (Cosmides and Tooby, 1996; Gigerenzer and Hoffrage, 1995), lend themselves easily to such simplification and have been shown to improve Bayesian reasoning. For instance, Bayesian responses to the mammography problem more than doubled when it was presented in natural-frequency format (Gigerenzer and Hoffrage, 1995). Although the theoretical bases of such improvements are debated (e.g., Barbey and Sloman, 2007, and continuing commentaries), most agree that substantial improvement in conformity to Bayes' theorem is achievable in this manner.

Bayesian reasoning also benefits from the use of visual representations of pertinent statistical information, such as Euler circles (Sloman et al., 2003) and frequency grids or trees (Sedlmeier and Gigerenzer, 2001), which further clarify nested-set relations. For instance, Figure 1 shows how the natural-frequency version of the mammography problem could be represented with a frequency tree to help individuals visualize the nested-set relations and how such information ought to be used to compute the posterior probability.

Observations

A remarkable feature of the standard approach to studying Bayesian reasoning is its inability to reveal how people *revise* their beliefs or subjective probabilities in light of newly acquired evidence. That is, in tasks such as the mammography problem, information acquisition is not staged across time (real or hypothetical), and researchers typically do not collect multiple "prior" and "posterior" (i.e., revised) probability assessments.

It is instead conveniently assumed that the base rate represents the subject's prior belief, $P(H)$, which the subject updates in light of "new" evidence, D. It is somewhat ironic that advocates of base-rate neglect have not noted (let alone warned) that, if people ignore base rates, it may be unwise to assume they represent the subject's prior probability. Would that not imply that the subject ignores his or her own prior probability?

Priors need not equal base rates, as many have noted (e.g., Cosmides and Tooby, 1996; de Finetti, 1964; Levi, 1983; Niiniluoto, 1981). The prior, $P(H)$, is in fact a *conditional* probability corresponding to one's personal probability of H, given all that they know prior to learning D (Edwards et al., 1963; de Finetti, 1972). In all real-life cases where no single, relevant base rate is ever explicitly provided, people may experience considerable uncertainty and difficulty in deciding precisely which base rate is the most relevant one to consider. For instance, imagine that the test result in the mammography problem is for a specific, real woman and not just an abstract one lacking in other characteristics. If her prior for H is contingent on the presence or absence of some of those characteristics, one could see how the base rate provided in the problem might be more or less relevant to the woman's particular case. If she has several characteristics known to elevate a woman's risk of breast cancer, then simply using the base rate for 40-year-old women as her prior would bias her revised assessment by leading her to underestimate the risk she faces. Conversely, she may have a configuration of characteristics that make her less likely than the average 40-year-old woman to develop breast cancer, in which case using the base rate as her prior would cause her to overestimate objective risk.

Clearly, the ideal base rate in such personal cases would be a sample of people who are just like the patient, yet since each of us is unique no such sample exists. In the absence of a single, ideal base rate, one must decide among a range of imperfect ones—a task involving decision under uncertainty. It might be sensible for the woman getting the screening to anchor on a relevant, available base rate, such as for women in her cohort, and then adjust it in light of other diagnostic characteristics that she knows she possesses. Yet, if people are overly optimistic (Taylor and Brown, 1988; Weinstein, 1989), we might anticipate systematic biases in adjustment, with underweighting of predisposing factors and overweighting of mitigating factors. This point about the possible role of motivated cognition also brings a key tenet of subjective Bayesianism to the fore—namely, that different individuals with access to the same information could have different degrees of belief in a given hypothesis, and they may be equally good Bayesians

as long as they are equally respectful of static and dynamic coherence requirements (Baratgin and Politzer, 2006).

Given that standard Bayesian reasoning tasks involve no assessment of a prior probability, they should be seen for what they are: conditional probability judgment tasks that require the combination of statistical information. When that information is fleshed out, it reveals the fours cells of a 2 × 2 contingency table, where $a = f(H \cap D)$, $b = f(H \cap \neg D)$, $c = f(\neg H \cap D)$, and $d = f(\neg H \cap \neg D)$. Going from left to right, the four boxes in the lowest level of the frequency tree in Figure 1 correspond to cells *a-d,* which have received much attention in the causal induction literature (Mandel and Lehman, 1998). We can restate Bayes' theorem as the following cell-frequency equalities, corresponding to short and long expressions given earlier, respectively:

$$P(H \mid D) = \frac{a}{a+c} = \frac{(a+b)/(a+b+c+d) \times a/(a+b)}{(a+b)/(a+b+c+d) \times a/(a+b) + (c+d)/(a+b+c+d) \times c/(c+d)}.$$

From this perspective, it is perhaps unsurprising why a greater proportion of subjects conform to Bayes theorem when they are given the frequencies *a-d* than when they are instead given the values equal to $(a + b)/(a + b + c + d)$, $a/(a + b)$, and $c/(c + d)$. That is, frequencies *a-c* support the easy computation of $a/(a + c)$. However, those improvements in performance, which pertain to static coherence constraints (Baratgin and Politzer, 2006), do not speak to other important facets of Bayesian reasoning, such adherence to dynamic coherence constraints, which are fundamental to Bayesian belief revision (Seidenfeld, 1979).

I do not intend for my observations to imply that the well-established findings I summarized earlier are incorrect. However, I believe greater care should be taken in labeling the type of performance measured in such experiments. "Statistical inference" would seem to be more appropriate than "Bayesian reasoning" given the limitations I have noted.

Future research on Bayesian reasoning would benefit from a richer conceptualization of what it is to "be Bayesian" and from better discussion of whether being non-Bayesian is necessarily irrational (Baratgin and Politzer, 2006; Lewis, 1976; Walliser and Zwirn, 2002). Future work would also benefit by breaking free of the typical methodological approach exemplified by the mammography problem. One avenue would be to collect prior and posterior assessments from subjects in experiments where information acquisition is staged (e.g., Girotto and Gonzalez, 2008), or where temporal staging is at least an important characteristic of the described problem, such as in the Monty Hall problem (Krauss and Wang, 2003) and Sleeping Beauty problem (Elga, 2000; Lewis, 2001). Another promising line involves assessing people's prior distributions for different types of real events (e.g., Griffiths and Tenenbaum, 2006).

The staging of information with repeated assessments was in fact a common methodological approach in Bayesian research prior to the 1970s, culminating in the classic work on conservatism by Ward Edwards and others (for a review, see Slovic and

Lichtenstein, 1971). Such approaches could be revisited in new forms and contrasted with other methods of information staging, such as the trial-by-trial information acquisition designs used in causal induction (e.g., Kao and Wasserman, 1993; Mandel and Vartanian, 2009) or category learning (e.g., Gluck and Bower, 1988; Shanks, 1990) studies.

For example, Williams and Mandel (2007) presented subjects with 28 problems prompting them for a conditional probability judgment. In each problem, subjects first saw 20 patient results presented serially. The subject saw whether the patient carried a virus hypothesized to cause a particular illness and whether the patient had the illness or not. Sample characteristics were varied so that $P(H|D)$ ranged from 0-1 over seven probability levels across the problems. Subjects exhibited a form of conservatism (cf. Edwards, 1968), overestimating low probabilities and underestimating high probabilities. The task illustrates the value of breaking free of the standard problem set. First, the trial-by-trial design better represents the information acquisition environment that ecological rationality theorists (e.g., Cosmides and Tooby, 1996; Gigerenzer and Hoffrage, 1995), have described as natural. That is, information acquisition in that task is more natural than in natural-frequency versions of standard problems because no statistical information is presented to the subject in written form. Rather, subjects learn about each case serially, more like they would have in the Paleolithic Era. Second, the design gets researchers away from studying average responses to a single problem with a unique data configuration. The authors would not have been able to detect conservatism if they had not explored problems for which the mathematical probabilities subjects were asked to judge covered the full probability range. Third, the induction paradigm, which presents information on cells *a*-*d* to subjects, easily lends itself to studying subjective cell importance, which can help take the cognitive processes subjects use to arrive at their judgments out of the proverbial black box. For instance, Williams and Mandel (2007) found that, when asked to assign subjective importance ratings to each of the fours cells, subjects assigned weight to irrelevant information, such as focusing on $\neg D$ cases when asked to judge $P(H|D)$, causing an underweighting of relevant information.

The issues I have raised, non-exhaustive as they are, draw attention to some important problems with the conventional approach to studying Bayesian reasoning in psychology that has been dominant since the 1970s. Rather than fostering pessimism, I hope my comments illustrate that there are good opportunities for future work to advance our understanding of how people revise or update their beliefs.

Acknowledgements

I thank Baruch Fischhoff, Gorka Navarrete, and two anonymous reviewers for helpful comments on earlier drafts of this paper.

References

Ayal, S., and Beyth-Marom, R. (2014). The effects of mental steps and compatibility on Bayesian reasoning. *Judgm. Decis. Mak*. 9, 226-242.

Baratgin, J., and Politzer, G. (2006). Is the mind Bayesian? The case for agnosticism.

Mind & Society 5, 1-38.
Barbey, A. K., and Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behav. Brain Sci.* 30, 241-297.
Cosmides, L., and Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cogn*. 58, 1-73.
Eddy, D. M. (1982). "Probabilistic reasoning in clinical medicine: Problems and opportunities," in *Judgment under Uncertainty: Heuristics and Biases*, ed. D. Kahneman, P. Slovic and A. Tversky (New York: Cambridge University Press), 249–67.
Edwards, W. (1968). "Conservatism in human information processing," in *Formal Representation of Human Judgment*, ed. B. Kleinmuntz (New York: Wiley), 17-52.
Edwards W., Lindman, H., and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychol. Rev.* 70, 193–242.
Elga, A. (2000). Self-locating belief and the sleeping Beauty problem. *Analysis* 60, 143-147.
Finetti, B. de (1964). "Foresight: its logical laws, its subjective sources," in *Studies in Subjective Probability*, ed. H. E. Kyburg and H. E. Smokler (New York, Wiley), 53–118 (1st ed., 1937).
Finetti, B. de (1972). "Probability, statistics and induction: their relationship according to the various points of view," in *Probability, Induction and Statistics. The Art of Guessing*, ed. B. de Finetti (London: Wiley), 141–228 (1st ed., 1959).
Gigerenzer, G., and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychol. Rev.* 102, 684–704.
Girotto, V., and Gonzalez, M. (2008). Children's understanding of posterior probability. *Cogn.* 106, 325-344.
Gluck, M. A., and Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *J. Exp. Psychol. Gen.* 117, 227-247.
Griffiths, T. L., and Tennenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychol. Sci.* 17, 767-773.
Hoffrage, U., Gigerenzer, G., Krauss, S., and Martignon, L. (2002). Representation facilitates reasoning: what natural frequencies are and what they are not. *Cogn*. 84, 343-352.
Kahneman, D., and Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cogn. Psychol*. 3, 430-454.
Kahneman, D., and Tversky, A. (1973). On the psychology of prediction. *Psychol. Rev.* 80, 237–251.
Kao, S.-F., and Wasserman, E. A. (1993). Assessment of an information integration account of contingency judgment with examination of subjective cell importance and method of information presentation. *J. Exp. Psychol. Learn. Mem. Cogn*. 19, 1363-1386.
Karvetski, C. W., Olson, K. C., Mandel, D. R., and Twardy, C. R. (2013). Probabilistic coherence weighting for optimizing expert forecasts. *Decis. Anal.* 10, 305-326.
Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative and methodologica l challenges. *Behav. Brain Sci.* 19, 1-53.
Krauss, S., and Wang, X. T. (2003). The psychology of the Monty Hall Problem:

Discovering psychological mechanisms for solving a tenacious brain teaser. *J. Exp. Psychol. Gen.* 132, 3-22.
Levi, I. (1983). Who commits the base rates fallacy. *Behav. Brain. Sci.* 6, 502–506.
Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *Philos. Rev.* LXXXV, 297–315.
Lewis, D. (2001). Sleeping Beauty: Reply to Elga. *Analysis* 61, 171-176.
Mandel, D. R. (2005). Are risk assessments of a terrorist attack coherent? *J. Exp. Psychol. Appl.* 11, 277-288.
Mandel, D. R., and Lehman, D. R. (1998). Integration of contingency information in judgments of cause, covariation, and probability. *J. Exp. Psychol. Gen.* 127, 269-285.
Mandel, D. R., and Vartanian, O. (2009). Weighting of contingency information in causal judgment: Evidence of hypothesis dependence and use of a positive-test strategy. *Q. J. Exp. Psychol.* 62, 2388-2408.
Niiniluoto, I. (1981). Cohen versus Bayesianism. *Behav. Brain. Sci*. 4, 349.
Sedlmeier, P., and Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *J. Exp. Psychol. Gen.* 130, 380–400.
Seidenfeld T. (1979). Why I am not an objective Bayesian: some reflections prompted by Rosenkrantz. *Theory Decis.* 11, 413-440.
Shanks, D. R. (1990). Connectionism and the learning of probabilistic concepts. *Q. J. Exp. Psychol.* 42A, 209-237.
Sloman, S. A., Over, D. E., Slovak, L., and Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organ. Behav. Hum. Decis. Process.* 91:296–309.
Slovic, P. and Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organ. Behav. Hum. Perform.* 6, 649-744.
Taylor, S. E., and Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychol. Bull.* 103, 193–210.
Villejoubert, G., and Mandel, D. R. (2002). The inverse fallacy: An account of deviations from Bayes's theorem and the additivity principle. *Mem. Cogn.* 30, 171-178.
Walliser B., and Zwirn, D. (2002). Can Bayes' rule be justified by cognitive rationality principles? *Theory Decis.* 53, 95–135.
Weinstein, N. D. (1989). Optimistic biases about personal risks. *Science* 264, 1232–1233.
Williams, J. J., and Mandel, D. R. (2007). "Do evaluation frames improve the quality of conditional probability judgment?," in *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*, ed. D. S. McNamara and J. G. Trafton (Mahwah, NJ: Erlbaum), 1653-1658.

Figure legend

Figure 1. Frequency tree and solution for the mammography problem.