'Confirmation bias' in rule discovery and the principle of maximum entropy

Edward Hoenkamp
Computer Science Department
University of California, Los Angeles

ABSTRACT

In scientific research as well as in everyday reasoning, people are prone to a 'confirmation bias', i.e. they tend to select tests that fit the theories or beliefs they already entertain. This tendency has been criticized by philosophers of science as not optimal. The behavior has been studied in a variety of psychological experiments on controlled, small-scale simulations of scientific research. Applying elementary information-theory to sequential testing during rule discovery, this paper shows that the biased strategy is not necessarily a bad one, moreover, that it reflects a healthy propensity of the subject (or researcher) to optimize the expected information on each trial.

INTRODUCTION

The standard scientific paper backs up the presented theory with corroborating evidence, and does not discuss at length findings that could falsify it. This does not reflect a dishonesty on the part of the scientist but primarily that she found what she was looking for: once a theory takes shape, the scientist is prone to look for evidence that confirms rather than disconfirms the predictions. It has long been observed that in scientific research as well as in everyday reasoning people tend to test cases that confirm their currently held hypotheses or beliefs. This tendency is called 'confirmation bias', but is actually an aggregate of several distinct phenomena. Aside from behavior during rule-discovery, it ranges from biased reasoning in inference tasks, such as Wason's four-card problem [Wason & Johnson-Laird, 1972], to bias in social perception as studied by Snyder and Swann [1978]. It not only turns up in learning situations, but it is also an important factor in the perseverance of beliefs after evidential discrediting [Ross, Lepper, Hubbard, 1975; Hoenkamp, 1987]. This paper uses the term to mean the strategies people use to discover a rule governing a set of data. A thorough analysis of the phenomenon from a Bayesian perspective can be found in [Klayman & Ha, 1987], which contains some fine points not mentioned in the present article. Another Bayesian approach is [Fischhoff & Beyth-Marom, 1983], which emphasizes shortcomings in hypothesis evaluation. Since philosophers of science such as Popper [1962] and Platt [1964] stress the importance of seeking disconfirmation, there seems to be a discrepancy between what people actually do, and what they ought to do. This paper uses a measure of information to assess conditions for the appropriateness of these competing strategies.

EXPERIMENTS ON RULE DISCOVERY

Wason [1960] designed an experiment to study how people behave when their beliefs about a rule are corroborated. Subjects were told that they had to guess a rule governing a number triple, and that 2-4-6 conformed to this rule. They were to figure out this rule by proposing other number triples. After each one the experimenter would tell whether it conformed to the rule. If the subject announced a rule, the experimenter would tell her whether it was correct. The rule the experimenter had in mind was 'increasing numbers'.

The following protocol typifies the trend of subjects to generate number triples expected to confirm their hypotheses [Wason & Johnson-Laird, 1972]. The '+' or '-' is used by the experimenter to indicate correctness.

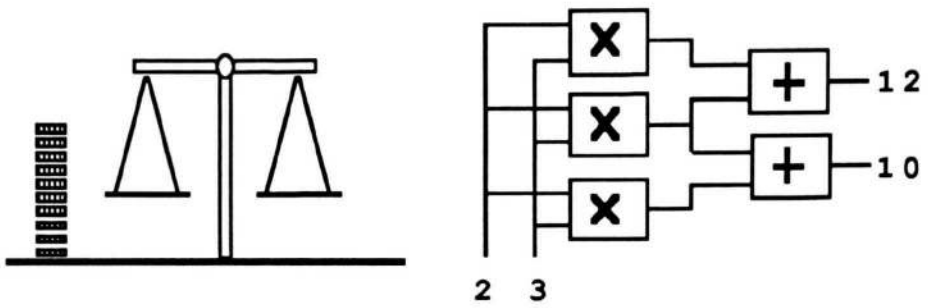
2-4-6 (+) given. 8-10-12 (+) two added each time. 14-16-18 (+) even numbers in order of magnitude. 20-22-24 (+) same reason. 1-3-5 (+) two added two preceding number. Announcement: 'starting with any number two is added each time' (incorrect). 2-6-10 (+) middle number is arithmetic mean of other two. 1-50-99 (+) same reason. Announcement: 'middle number is the arithmetic mean' (incorrect). 3-10-17 (+) same number, seven, added each time. 0-3-6 (+) three added each time. Announcement: 'the difference between two numbers next to each other is the same' (incorrect). 12-8-4 (-) same number subtracted each time. Announcement: 'adding a number, always the same one' (incorrect). 1-4-9 (+) any three numbers in order of magnitude. Announcement: 'any three numbers in order of magnitude' (correct).

This experiment has been replicated many times with simple to very intricate variations. Some researchers tried to induce a tendency to falsify [Mynatt et al., 1977; Tweney et al., 1982], or used a broader rule [Gorman & Gorman; 1984]. Others worked with groups [Gorman, Gorman, Latta, Cunningham, 1984]. Domains other than number triples have been used; Mynatt et al. [1977] simulated a universe on a computer screen, where subjects had to find the rules that governed the deflection of particles near objects. Kern [1982] used an imaginary planet, and subjects had to locate the place where creatures landing on the planet would stay alive. Time and time again the tendency to confirm showed up. And it seems representative for scientists' actual behavior, as was found by Mitroff [1974] who interviewed scientists at NASA before the first Apollo moon landing. Given that this proclivity is so pervasive, the question is: are people really doing it the wrong way? To find an answer we have to investigate if, or under what circumstances, there exists an optimal test strategy.

CHOOSING AN OPTIMAL TEST STRATEGY

There are at least two ways in which a strategy for sequential testing could be optimal [Tweney et al., 1981]. To decide which of two strategies is optimal, one could take the one that best complies with an established criterion for rationality, such as Popper's falsification principle. Or one could opt for the most efficient one (in terms of time, money etc.). The two criteria are independent, but if the most efficient strategy turns out the more successful on average, the established criterion is immaterial. But how can we quantify the efficiency of a strategy? I define a strategy to be the most efficient if the expected cost of testing is minimal. If every test costs the same, our goal would be to minimize the number of tests. Consequently, we want each test to be as informative as possible. But wait, informative as possible for what purpose?

Suppose a subject in Wason's task proposes a triple 10-12-14 to test 'subsequent even numbers'. Wason sees this as a confirmatory strategy. The concept is confusing however: the triple confirms the subject's rule, but it could disconfirm the experimenter's rule. In his criticism of Wason, Wetherick [1962] uses the term *positive test* for this case. A triple 1-3-5 to test 'subsequent even numbers' would then be a negative test, which in contrast could be confirmed by the experimenter. I will call a positive test strategy one that relies on positive testing to investigate a hypothesis (analogously for 'negative test strategy'). Klayman and Ha [1987], using the same distinction, investigate which strategy has the highest probability of falsifying the hypothesis. This is much in line with Popper's [1962] principle to strive for falsification: accumulating confirmatory evidence cannot prove a theory correct, but one falsification is enough to show the theory is incorrect. However, as other philosophers have shown [Kuhn, 1970; Feyerabend, 1975], scientists will not abandon their theories, but instead make changes to their theories that will account for the new findings. To understand this behavior, we should not look for test strategies that are most likely to show a theory is flawed. Instead, we should look for one that shows where the flaws are located, so as to make optimal changes. I will stay with the concepts of positive and negative test strategies, but compare them not by their probability to falsify, but by the information they provide per test.



(a) (b)

Figure 1. Two examples of sequential testing. **a.** Nine coins, one of which is lighter or heavier than the others. Determine this one in the least number of weighings. **b.** For the given three multipliers and two adders, make the minimum number of measurements to determine the malfunctioning component(s).

INFORMATION CONTENT OF A TEST

To measure the information content of a test, I will borrow concepts from information theory. We denote the possible outcomes a_i of an experiment with their probabilities p_i as the finite scheme A:

$$\begin{bmatrix} a_1 & a_2 & \dots & a_n \\ p_1 & p_2 & \dots & p_n \end{bmatrix}, \text{ e.g. for a 'true die' the scheme would be } \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{bmatrix}.$$

The information contained in such a scheme is called entropy. For the finite scheme A above, the entropy is defined as:

$$\mathcal{H}(p_1, p_2, \dots, p_n) = -\sum_{i=1}^{n} p_i * \log(p_i) \qquad (x1)$$

The entropy can be viewed as the uncertainty taken away after the outcome of the experiment becomes known. Consequently, if one of the probabilities is 1, the entropy is 0, since in that case the outcome is certain. Now, if a choice can be made among several different schemes, the one with maximum entropy reduces the most uncertainty. Figure 1 shows two applications of this idea to sequential testing. For example, in figure 1a the choice is among the numbers of coins to be put in each pan. (For example, to maximize the entropy on the first step, one has to put 3 coins in each pan). Another example is the proposal by De Kleer and Williams [1987] for diagnosing multiple faults in electronic circuits (see figure 1b).

The concrete examples above can be generalized to optimizing test strategies in general. Suppose a scientist (or a subject) generates an hypothesis to describe phenomena in some domain D. Let us denote the intended phenomena as T (target), and the set described by the hypothesis as H. Figure 2 shows the four possible locations of a new observation. The figure depicts the general case, but the reader can think of the 2-4-6 task as an example.

The black arrows show that the new observation may end up in either $H \cap T$, in which case the hypothesis is corroborated, or in $H \cap T^c$, and then the hypothesis is falsified. The positive and negative strategies can be represented as the following schemes for some p^+ and p^-:

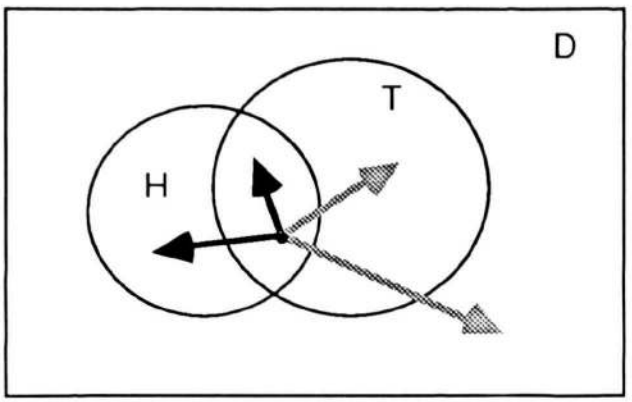


Figure 2. The places where the new observation may lie. H and T are the hypothesized set and the target set in domain D. The black and grey arrows indicate the positive and negative test strategies respectively.

$$S^+ = \begin{bmatrix} H \cap T & H \cap T^c \\ p^+ & 1-p^+ \end{bmatrix} \quad \text{and} \quad S^- = \begin{bmatrix} H^c \cap T & H^c \cap T^c \\ p^- & 1-p^- \end{bmatrix}$$

The p^+ and p^- can thus be written as the conditional probabilities $p^+ = P(T/H)$, and $p^- = P(T/H^c)$, with H^c to denote the complement of H. In comparing both strategies, a few qualitative remarks can be made. First, a scheme may contain no information at all, namely if one of the probabilities equals 1 (since then the outcome is certain). This is the case if p^+ or p^- equal 1 or 0. But note, there is an overlap between H and T containing the element(s) of T used to formulate H in the first place. This leaves $p^-=0$ and $p^+=1$ to consider. If $p^-=0$ then H includes T. In other words if the hypothesis is too general, a positive strategy is more efficient than a negative. For $p^+=1$, T contains H, and so in that case a negative strategy is better (we will come back to this).

A second observation is that both strategies contain the same maximum information (of 1 bit). $\mathcal{H}(S^-)$ depends on p^-, and thus on the size of the domain, whereas $\mathcal{H}(S^+)$ depends on the overlap of H and T only . All in all, it may be that the positive strategy is not as bad as it might have looked. How good or bad it is quantitatively, will be discussed in the next section.

COMPARING THE INFORMATION CONTENT OF BOTH STRATEGIES

To investigate under precisely which circumstances a positive strategy is the more efficient, i.e. produces the most information per test, the following inequality can be solved:

$$\mathcal{H}(S^+) > \mathcal{H}(S^-)$$

according to definition (x1) this means that

$$-(p^+ * \log(p^+) + (1-p^+) * \log(1-p^+)) > -(p^- * \log(p^-) + (1-p^-) * \log(1-p^-))$$

Using the property that $p*\log p + (1-p)*\log(1-p)$ is a convex function on [0,1] symmetric around .5 this simplifies to $p^- < p^+ < 1-p^-$, or equivalently

$$\frac{p^+}{p^-} > 1 \qquad (x2)$$

and

$$p^+ + p^- < 1 \qquad (x3)$$

Let's take a closer look at (x2). According to Bayes' rule, updating the probability of a hypothesis H upon receiving datum T satisfies

$$\frac{P(H/T)}{P(H^c/T)} = \frac{P(T/H)}{P(T/H^c)} * \frac{P(H)}{P(H^c)}$$

Writing out the fraction in formula (x2) as $P(T/H)/P(T/H^c)$ shows it is the likelihood ratio (the second term) in Bayes' rule. If this ratio is greater then 1, the datum is called diagnostic. So to rephrase formula (x2): for a positive strategy to be optimal, the target elements must be diagnostic for the hypothesis. At the end of the paper I will return to the relative merits of both conditions.

Noisy data. The experiments that make up most of the literature, are based on error-free feedback. Outside the laboratory the situation is often far from that ideal. Does this have an influence on which strategy should be preferred? In the rule-discovery tasks discussed here it means that the subject receives a 'correct' where an 'incorrect' would be in place and vice versa. In the presence of error, a strategy has to be chosen that maximizes information per test *on average*. In terms of information-theory what we have is a noisy channel of a particular kind (a binary symmetric channel) over which the feedback is sent. Given that enough tests can be performed and that the error-rate is less than .5, the actual feedback can be recovered (using an optimal coding scheme). Interestingly enough, under these circumstances the scheme that maximizes the information in the error-free case also maximizes the average information in the presence of error. It follows that the inequalities (x2) and (x3) remain valid.

It should be noted that formula (x1) can be derived from three very simple and plausible axioms [Khinchin, 1957], such as that adding impossible events to the scheme doesn't change the entropy[1]. So the results in this paper depend only on the acceptance of those axioms, and the cost function for sequential testing. Yet, as the next section will show, several interesting psychological findings can be easily understood this way.

APPLICATIONS OF THE THEORY TO EXPERIMENTAL FINDINGS

Recall that we are talking about tasks in which successive tests are needed for a discovery, and that reducing the number or the cost of tests is achieved by choosing a strategy that maximizes the entropy on each trial. We shall now see how an assorted sample of observations can be explained in this manner[2].

Wason's 2-4-6 induction task. As we saw before, in this task, the subject starts with a hypothesis (subsequent even numbers) that is a subset of the experimenter's rule (increasing numbers). In this case $P(T/H)=1=p^+$, so that inequality (x3) is not satisfied, and therefore a negative strategy is be preferable. This is exactly what the experiment showed.

The 'first confirm later disconfirm' strategy [e.g. Gorman & Gorman, 1984]. In rule-discovery tasks, the successful subject usually starts with a positive strategy, and later shifts more to a negative strategy. Suppose a set C of observations has been confirmed for hypoth-

---

[1]The other two are: the entropy is maximal if all probabilities are equal, and the entropy of two schemes equals the entropy of the first plus the expectation of the second given the outcome of the first.

[2]Which only shows how difficult it is to exorcise confirmation bias.

esis H. If the tester is not simply replicating observations, the p^+ has then decreased[1], namely from the initial P(T/H) to P(T/H-C), while p^- remains the same. After a while inequality (x2) will not be fulfilled any longer at which point the tester should switch to a negative strategy.

The 'win-stay, lose-shift' strategy. Studies in concept identification have shown that once learners have a hypothesis about reinforced responses, they will stick to that hypothesis even if later other responses are also reinforced. The hypothesis is changed only if falsifying information is encountered [Trabasso & Bower, 1968]. The rule cannot be true in general. The maximum entropy for S^+ occurs for $p^+=.5$. And as we have seen just before, p^+ decreases for an S^+, so the suggestion is justified if $p^+ > .5$, i.e. if more than half the hypothesis set is in the target. Indeed, this restriction holds in the cases discussed by Trabasso and Bower [1968].

Positive strategies work better for groups. Condition (x2) states that for a positive strategy to work the target elements must be diagnostic for the hypothesis. There is considerable literature about people's neglect in using the likelihood ratio in evaluating probabilities [Kahneman, Slovic, Tverski, 1982]. However, Trope and Bassok [1982] showed that diagnosticity is a major determinant in people's preference for a particular information-gathering strategy. That is, if given the opportunity to compare, they opt for the hypothesis with the highest diagnosticity. So one can expect if a group of people, such as a scientific team, generates various hypotheses, the one with highest diagnosticity will be recognized. In that case it seems probable that one satisfying (x2) will occur, and therefore will prevail. Indeed, groups using positive strategies are better in a rule-discovery task[2] than individuals [Gorman, Gorman, Latta, Cunningham, 1984].

A negative strategy doesn't work if the rule is too general. If T grows to cover a larger part of D, $\mathcal{H}(S^-)$ decreases. *In other words, if the rule becomes more general the negative* strategy will give less and less information per trial. *This may explain the finding on a* variation of the 2-4-6 task. *Gorman and Gorman [1984] used two more general rules besides* the 'ascending numbers', namely 'at least one even number' and *'no two numbers can be the* same'. They indeed found that even subjects who were encouraged to use S^- were not successful in discovering the rule.

Feedback in the presence of noise. In most experiments the feedback is error-free. But as mentioned before, conditions (x2) and (x3) remain valid in the presence of error, if the subject (or researcher) is allowed to perform many tests. Two things change subject's behavior, however. First, the information per test is lower, so more tests have to be performed. Second, an optimal coding scheme requires that tests have to be replicated. Given that a positive strategy is appropriate, this should induce long stretches of positive tests. Kern [1982] asked subjects to partake in a computer simulation where creatures had to be placed on an imaginary planet. They had to discover a line on one of which sides the creatures died. She found that if the feedback was random on a proportion of the trials, a strong positive testing tendency ensued. A strong tendency to replicate was found in [Gorman, 1986], confirming the need for replications in the face of noise.

Klayman and Ha's approach. In the Bayesian approach taken by Klayman and Ha [1987], the preferred strategy is the one most likely to falsify the hypothesized prediction. In their formalization this means that the positive strategy is preferable precisely if $P(T^c/H) > P(T/H^c)$,

---

[1]Except for the degenerate case where $p_c=1$.

[2]The task was 'Eleusis' in which a rule governing the appearance of playing cards had to be discovered.

i.e. 1-P(T/H) > P(T/H^c), which is equivalent to inequality (x3). Klayman and Ha show that the inequality holds under realistic circumstances. In other words, people's positive strategy is very often appropriate. Their approach, however, misses inequality (x2), and thus leaves unexplained the phenomena mentioned above that depend on it. In addition, (x3) didn't have to be postulated, it follows automatically under the plausible assumption that a good strategy optimizes the cost of sequential testing. It would be interesting to design an experiment where (x3) holds, and (x2) doesn't. The prediction is that a positive strategy would give the highest probability for falsification, whereas a negative one would produce the best information to change the theory.

CONCLUSION

This paper compared people's actual behavior in a rule discovery tasks with the behavior they should exhibit according to some canon of rationality. It did so by describing people's discovery behavior as sequential testing for which the total cost of the trials has to be optimized. The paper showed that under that criterion, and given realistic circumstances, a positive strategy is often the best one. The derived conditions were shown to explain the degree to which people are successful in rule-discovery tasks in a spectrum of experimental settings. At the same time they may suggest variations on the task (such as changing the diagnosticity of the target).

Little attention has been paid in the psychological literature to other strategies of inquiry (but see [Tukey, 1986] for an exception). The next step in this research therefore is to analyze the relative merits of such strategies in the way it was done in the present paper for positive and negative strategies. If this also leads to the formulation of new conditions (analogous to (x2) and (x3)), this may suggest experiments that shed additional light on people's modes of inquiry.

Acknowledgements

This work was supported by a grant from the Netherlands Organization for Scientific Research (NWO), during a sabbatical leave from NICI, Nijmegen, the Netherlands. I'm grateful to Hector Geffner, Charles Wharton, Vicky Breckwich, Claudia Lange, and Trent Lange for comments on an earlier version of the paper.

REFERENCES

De Kleer, J. & Williams, B. (1987). Diagnosing multiple faults. *Artificial Intelligence,* 32, 97-130.

Fischhoff, B. & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review,* 90, 239-260.

Gorman, M. (1986). How the possibility of error affects falsification on a task that models scientific problem solving. *British Journal of Psychology,* 77, 85-96.

Gorman, M. & Gorman, M. (1984). A comparison of disconfirmatory, confirmatory and control strategies on Wason's 2-4-6 task. *The Quarterly Journal of Experimental Psychology,* 36A, 629-648.

Gorman, M., Gorman, M., Latta, R., Cunningham, G. (1984). How disconfirmatory, confirmatory and combined strategies affect group problem solving. *British Journal of Psychology,* 75, 65-79.

Hoenkamp, E. (1987). An analysis of psychological experiments on non-monotonic reasoning. *Proceedings of IJCAI-87,* 115-118.

Kahneman, D., Slovic, P., & Tverski, A. (1982). Judgment and uncertainty: Heuristics and biases. New York: Cambridge UP.

Kern, L. (1982). The effect of data error in inducing confirmatory inference strategies in scientific hypothesis testing. Unpublished PhD thesis. Ohio State University.

Khinchin, A. (1957). *Mathematical foundations of information theory*. New York: Dover.

Klayman, J. & Ha, Y-W. (1987). Confirmation, disconfirmation, and hypothesis testing. *Psychological Review*, 94, 2, 211-228.

Kuhn, T. (1970). *The structure of scientific revolutions*. (2nd edition). Chicago: University of Chicago Press.

Feyerabend, P. (1975). *Against method*. London: Verso Editions.

Mitroff, I. (1974). Norms and counter-norms in a select group of the Apollo moon scientists: A case study of the ambivalence of scientists. *American Sociological Review*, 39, 579-595.

Mynatt, C., Doherty, M. & Tweney, R. (1978). Consequences of confirmation and disconfirmation on a simulated research environment. *The Quarterly Journal of Experimental Psychology*, 30, 395-406.

Popper, K. (1962). *Conjectures and refutations*. New York: Basic Books.

Platt, J. (1964). Strong inference. *Science*, 146, 347-353.

Ross, L., Lepper, M. & Hubbard, M. (1975). Perseverance in self-perception and social perception: Biased attributional processes in the debriefing paradigm. *Journal of Personality and social psychology*, 32, 880-892.

Snyder, M. & Swann, W. (1978). Hypothesis-testing processes in social interaction. *Journal of Personality and social psychology*, 36, 1202-1212.

Trope, Y. & Bassok, M. (1982). Confirmatory and diagnosing strategies in social information gathering. *Journal of personality and social psychology*, 43, 22-34.

Tukey, D. (1986). A philosophical and empirical analysis of subjects' modes of inquiry in Wason's 2-4-6 task. *Quarterly Journal of Experimental Psychology*, 38A, 5-33.

Tweney, R., Doherty, M., Worner, W., Pliske, D., Mynatt, C. (1980). Strategies of rule discovery in an inference task. *Quarterly Journal of Experimental Psychology*, 32, 109-123.

Tweney, R., Doherty, M., Mynatt, C. (1981). *On scientific thinking*. New York: Columbia UP. Introduction to chapter IV.

Trabasso, T. & Bower, G. (1968). *Attention in learning*. New York: Wiley.

Wason, P. (1960). On the failure to eliminate hypotheses in a conceptual task. *The Quarterly Journal of Experimental Psychology*, 12, 129-140.

Wason, P. & Johnson-Laird, P. (1972). *Psychology of reasoning: Structure and content*. Cambridge: Harvard UP.