Cognition, 32 (1989) 1–24

Intuitive reasoning about probability: Theoretical and experimental analyses of the "problem of three prisoners"*

SHINSUKE SHIMOJO

Nagoya University School of Medicine, Japan, and The Smith-Kettlewell Eye Research Institute, San Francisco

SHIN'ICHI ICHIKAWA

Department of Education, Tokyo Institute of Technology, Japan

Received June 1987, final revision accepted September 1988

Abstract

Shimojo, S., and Ichikawa, S. 1989. Intuitive reasoning about probability: Theoretical and experimental analyses of the "problem of three prisoners". Cognition, 32: 1–24.

Among various Bayesian problems of probability, the "problem of three prisoners" (Lindley, 1971; Mosteller, 1965) is an especially good example which illustrates the drastic discrepancy between intuitive reasoning and mathematical formal reasoning about probability. In particular, it raises intriguing questions concerning the mathematical and cognitive relevance of factors such as prior probabilities and the context in which certain information is given. In the current paper, we report a new version of the problem which turned out to be even more counterintuitive. This new version was also designed so that different inferential schemes would lead to separate estimates of posterior probability. The data obtained from questionnaires and theoretical analyses of the original and modified problems suggest that: (1) The psychological processes of intuitive reasoning are qualitatively different from mathematical reasoning. (2) The

*The authors would like to thank Naoki Abe of University of Pennsylvania, Kazuo Shigemasu of Tokyo Institute of Technology, Hiroshige Takeichi of University of Tokyo, Shin'ichi Mayekawa of the National Center for University Entrance Examination for their insightful suggestions, and Lawrence M. Parsons of Massachusetts Institute of Technology and David Jones of Stanford University for their comments on the early draft. Shinsuke Shimojo is supported by the Fellowships of the Japan Society for the Promotion of Science for Japanese junior scientists. This study is partly supported by Grant-in-Aid for Scientific Research (No. 6176036), Ministry of Education, Science, and Culture of Japan, to Shin'ichi Ichikawa. Reprint requests should be sent to Shinsuke Shimojo, Department of Psychology, University of Tokyo, Komaba, Meguro-ku, Tokyo, Japan 153.

tendency to neglect prior probabilities (Tversky & Kahneman, 1974, 1982) is not always the critical factor for illusory judgments. (3) Intuitive judgments can be categorized by several, distinctive propositional beliefs from which the judgments are apparently derived. We call these prototypical, crude beliefs "subjective theorems," and discuss their nature and roles in the current paper.

Introduction

The problem of three prisoners

> *Problem 1:* Three men, A, B and C, were in jail. A knew that one of them was to be set free and the other two were to be executed. But he didn't know who was the one to be spared. To the jailer who did know, A said, "Since two out of the three will be executed, it is certain that either B or C will be, at least. You will give me no information about my own chances if you give me the name of one man, B or C, who is going to be executed." Accepting this argument after some thinking, the jailer said "B will be executed." Thereupon A felt happier because now either he or C would go free, so his chance had increased from 1/3 to 1/2. This prisoner's happiness may or may not be reasonable. What do you think?

This is the original version of the "problem of three prisoners" (partially revised from Mosteller, 1965, and Lindley, 1971). Even though A's first argument that the information is irrelevant to A's chances of survival sounds reasonable, it is still very difficult for most of us to resist the intuitive feeling that his chance should have increased with this information. According to the Bayesian analysis of the problem, A's chances in fact do not change in this case (Bar-Hillel & Falk, 1982; Lindley, 1971; Mosteller, 1965). The chances of survival for each of the three men prior to the conversation between A and the jailer, that is, P(A), P(B) and P(C), should be equal because no specific information indicated otherwise (the "principle of insufficient reason" or the Bayes' axiom). Thus,

$$P(A) = P(B) = P(C) = 1/3. \quad (1)$$

Let us now consider the probability of the jailer answering "B will be executed" in each of the three cases. Assuming that: (a) the jailer always tells the truth, and (b) he has no preference between naming "B" and "C" if both are to be executed, it is reasonable to assign

$$P(b|A) = 1/2, \quad P(b|B) = 0, \quad P(b|C) = 1, \quad (2)$$

where b denotes the datum (the jailer's answer "B will be executed"), and A, B and C respectively denote each one's survival. Thus, P(b|A) is the conditional probability of b given that it is known that A will be freed. Note that the jailer would have a 50–50 choice in his answer between "B" and "C" if A is to be freed, whereas there is no such choice under the condition that C is to be freed (and b is impossible under the condition that B is to be freed).

According to Bayes' theorem, the probability that A will be freed given the jailer's answer that B will be executed, P(A|*b*), is expressed in the form:

$$P(A|b) = \frac{P(b|A)P(A)}{P(b|A)P(A) + P(b|B)P(B) + P(b|C)P(C)}. \quad (3)$$

Substituting Equations (1) and (2) into (3),

$$P(A|b) = \frac{(1/2)\cdot(1/3)}{(1/2)\cdot(1/3) + 0\cdot(1/3) + 1\cdot(1/3)} = 1/3.$$

Thus, Bayes' theorem defines P(A|b) as the ratio of probability P(b|A)P(A) to the simple sum of all the conceivable probabilities of b, P(b|A)P(A) + P(b|B)P(B) + P(b|C)P(C) (see Figure 1a for illustration). The implication is apparently that the information on the other's fate is simply irrelevant to one's own chance (the "irrelevant, therefore invariant" notion). Even though some people might take this as the lesson from this particular problem,[1] it is not a rule generally applicable to this type of Bayesian problem, as demonstrated in the modified version of the problem:

Problem 2: Three men, A, B and C, were in jail. One of them was to be set free and the other two were to be executed. A had reason to believe that their chances of being freed were: A: 1/4, B: 1/4, C: 1/2. After their fates had been decided, A, who didn't know the outcome of the decision, asked the jailer, who did. "Since two out of the three will be executed, it is certain that either B or C will be, at least. You will give me no information about my own chances if you give me the name of one man, B or C, who is going to be executed." Accepting this argument, the jailer said, "B will be executed." Thereupon A felt happier because now either he or C would go free, so his chance had increased from 1/4 to 1/2. This prisoner's happiness may or may not be reasonable. What do you think?

[1] "... and mathematics comes round to commonsense after all." (Mosteller, 1965).

Figure 1. *A diagram for Bayesian scheme of estimating P(A|b). (a): Problem 1. (b): Problem 2 (see Introduction for the problems). The left-most column lists three possible cases, the second column lists the probability of each case. The third column then lists the probability of event b (the jailer answering "B will be executed") in each case. Finally, the right-most column shows the joint probability of event b for these cases. As shown in the bottom equation, the probability of A's survival after the jailer's answer, P(A|b), is defined as a ratio of the probability of "A survives* and *the event b occurs" (P(b|A)P(A)) to the sum of all joint probabilities of event b.*

(a)

	Three possible cases	Probability of jailer's saying "B will be executed" in each case			Joint probability
	↓		↓		
(I)	A: set free B: executed C: executed	P(A) = 1/3	P(b|A) = 1/2	→	P(b|A)P(A) = 1/6
(II)	A: executed B: set free C: executed	P(B) = 1/3	P(b|B) = 0	→	P(b|B)P(B) = 0
(III)	A: executed B: executed C: set free	P(C) = 1/3	P(b|C) = 1	→	P(b|C)P(C) = 1/3

$$P(A|b) = \frac{P(b|A)P(A)}{P(b|A)P(A) + P(b|B)P(B) + P(b|C)P(C)}$$

$$= \frac{(1/6)}{(1/6) + 0 + (1/3)} = 1/3$$

(b)

	Three possible cases	Probability of jailer's saying "B will be executed" in each case			Joint probability
	↓		↓		
(I)	A: set free B: executed C: executed	P(A) = 1/4	P(b|A) = 1/2	→	P(b|A)P(A) = 1/8
(II)	A: executed B: set free C: executed	P(B) = 1/4	P(b|B) = 0	→	P(b|B)P(B) = 0
(III)	A: executed B: executed C: set free	P(C) = 1/2	P(b|C) = 1	→	P(b|C)P(C) = 1/2

$$P(A|b) = \frac{P(b|A)P(A)}{P(b|A)P(A) + P(b|B)P(B) + P(b|C)P(C)}$$

$$= \frac{(1/8)}{(1/8) + 0 + (1/2)} = 1/5$$

In this new version of the problem, the prior probabilities are explicitly stated as

$$P(A) = 1/4, \quad P(B) = 1/4, \quad P(C) = 1/2. \tag{4}$$

Again, we assume that: (a) the jailer always tells the truth, and (b) he has no preference between naming "B" and "C" if both are to be executed:

$$P(b|A) = 1/2, \quad P(b|B) = 0, \quad P(b|C) = 1. \tag{5}$$

Substituting Equations (4) and (5) into (3), we obtain

$$P(A|b) = \frac{(1/2)\cdot(1/4)}{(1/2)\cdot(1/4) + 0\cdot(1/4) + 1\cdot(1/2)} = 1/5.$$

Thus, the Bayesian estimate of P(A|b) is in fact smaller than the prior probability P(A) (see Figure 1b for illustration; see also Appendix 1 for the gen-

eral mathematical condition for $P(A|b) < P(A)$). This is even more counterintuitive, particularly from A's viewpoint, because at least one of the rivals is no longer a candidate for release. (Our assumption about $P(b|A)$ may be controversial: see Experiment 3 and General Discussion.)

The Bayesian solutions of these problems appear very counterintuitive, even to those who are familiar with Bayes' theorem. It suggests that there is a line of intuitive reasoning functionally independent of formal mathematical reasoning. The Bayesian solution for the modified version of the problem implies that even the "irrelevant, therefore invariant" notion, which was the other intuitively acceptable way to answer the original version, is actually inappropriate. The goal of the current study is to understand the distinctive nature of intuitive reasoning about probability, and the relationship between processes of intuitive reasoning and of formal, mathematical reasoning. More specifically, we will examine some beliefs which are underlying people's intuitive reasoning, and discuss why the Bayesian solution to the problem is so counterintuitive.

Pilot study

Before designing the experiments, we intensively interviewed four graduate student subjects. All of them had statistical backgrounds equivalent to or better than the level of an introductory course, but none had detailed knowledge of Bayesian statistics. They were given Mosteller–Lindley's original version and our modified version of the problem (Problems 1 and 2), and asked to verbalize their processes of thinking as much as possible, although they were allowed to use a pen and a sheet of paper. No time constraint was imposed.

The results suggested that they actually employed one of several inferences such as: "Now there are only two cases possible (A or C to be freed), so the chance for A should be 1/2"; "The ratio of posterior probabilities should remain the same as the ratio of prior probabilities. Since two men remain, and the ratio of their chances was 1:1 (1:2 in Problem 2), the posterior chance for A should be 1/2 (1/3)"; or "The information is irrelevant, so the chance should not change from 1/3 (1/4)." These inferences were apparently based on general beliefs concerning the nature of probability. We refer to such beliefs as "subjective theorems" because they seem to be expressed best in a propositional or a declarative form. In particular, the above-mentioned three prototypes correspond to the following three subjective theorems:

"Number of cases" theorem: When the number of possible alternatives is N, the probability of each alternative is $1/N$.

Table 1. *The subjective theorems and estimates of P(A|b) based on these theorems*

Theorems	Problem 1	Problem 2
Bayes' theorem	1/3	1/5
"Number of cases"	1/2	1/2
"Constant ratio"	1/2	1/3
"Irrelevant, therefore invariant"	1/3	1/4

"Constant ratio" theorem: When one alternative is eliminated, the ratio of probabilities for the remaining alternatives is the same as the ratio of prior probabilities for them.

"Irrelevant, therefore invariant" theorem: If it is certain that at least one of the several alternatives (A_1, A_2, ..., A_k) will be eliminated, and the information specifying which alternative to be eliminated is given, it does not change the probability of the other alternatives (A_{k+1}, A_{k+2}, ..., A_N).

These theorems may be considered as "heuristics" (Tversky & Kahneman, 1974), but to emphasize their apparently declarative nature, we stick to "subjective theorems" in the current paper.[2] Table 1 summarizes the subjective theorems and estimates of P(A|*b*) based on these theorems for each problem.

The subjects occasionally shifted from one to another subjective theorem, or wavered between two incompatible subjective theorems. Even after they reported that they could "follow and understand" the mathematically correct reasoning, they still felt that it was against their intuition. The authors' anecdotal experience of discussing these problems with people, including experts in mathematics and statistics, agreed with the last point, particularly for the modified version. (This is reminiscent of various visual illusions in which the observer's knowledge about the physical stimulus does not much help to avoid illusory perceptual judgment.) Thus, we seem to have found an even more counterintuitive version of the "problem of three prisoners."

In formal mathematical reasoning, solutions can be "correct" or "incorrect," but never be "illusory." And yet, the above-mentioned observations raise a possibility that there may be a mental module of intuitive reasoning which can be more or less independent of formal mathematical reasoning.

The purpose of the following experiments is to investigate how people intuitively solve the problem and to understand the cognitive relationship between their intuitive solutions and Bayesian solutions.

[2]We say "apparently declarative" only because the subjects' protocols could easily be categorized into several types, each of which could be stated declaratively, almost as if it had been a proposition. However, we do recognize that a proposition should be translated into a procedural algorithm to be executable (see General Discussion for further discussion).

Experiments

General method

All three experiments used questionnaires administered by group. Each questionnaire consisted of two or three of seven variations of the "three prisoners' problem" (see Appendix 2). The order and organization of these problems in each experiment is shown in Table 2.

Subjects

One hundred and sixty-one students (11 females, 145 males, 5 anonymous) of the University of Tokyo participated in three experiments. They were freshmen and sophomores in the College of Liberal Arts and General Sciences. Very few of them had taken an introductory course in statistics, although as college students they had a reasonable background in mathematics. They were not familiar with Bayes' theorem and terminology, and none participated in more than one experiment. The number of subjects participating in each experiment is listed in Table 2.

Questionnaires and instructions

The subjects were divided into two subgroups and given the same problems, but in different sequences in each experiment (see Table 2 and Appendix 2).

All the subjects were given the following instructions before they started reading the questionnaire:

> This is not a quiz or an examination in mathematics, but a psychological questionnaire concerning intuitive reasoning on probability. So, please

Table 2. *Designs of the experiments (see Appendix 2 for the problems)*

			Problems on		
Exp.	Group	Subjects	Page 1	2	3
1	1	22	1	2	1
	2	19	2	1	2
2	1	25	3	4	3
	2	29	4	3	4
3	1	15	5	6	–
	2	15	6	5	–

> show your reasoning for the problems in a naive, intuitive and immediate fashion. Use formulas and equations if—and only if—they occur to you intuitively. Do not return to a previous page to change your answer. You will have three problems (in Experiments 1 and 2; two problems in Experiment 3), and be given a maximum of 5 minutes for each.

Experiment 1

The protocols in the pilot study indicated that the Bayesian estimates of the prisoners' probability are difficult to accept, even for those who can "understand" the logic of the formal solution. Further, the modified version of the problem seems to be even more counterintuitive. It was important to determine how naive subjects rated the probability for these two problems. The pilot study suggested that people tend to choose one of several theorems. If so, their estimates of the probability should be one of a few values. This prediction was tested in Experiment 1.

In the original problem (Problem 1), the "irrelevant, therefore invariant" and the Bayesian theorems give the same estimate ($P(A|b) = 1/3$), and the "number of cases" and the "constant ratio" theorems give the same estimate of probability ($P(A|b) = 1/2$: see Table 1). However, these inferential schemes, including the Bayesian, give different values of probability in the modified version of the problem (Problem 2). Thus, a more objective categorization of the subjects' reasoning may be expected with the modified version.

The second issue addressed in Experiment 1 is as follows. The major difference between the original and the modified versions of the problem was that the prior probabilities (1/4, 1/4, 1/2) are explicitly stated in the modified version, whereas they are just implicitly assumed to be equal (1/3, 1/3, 1/3) in the original version. The inattention to the prior probabilities is a prominent tendency of intuitive judgment under uncertainty (Tversky & Kahneman, 1974, 1982). Thus, exposure to the modified version may shift subjects' attention to the prior probabilities, and consequently change later estimates of probability for the original version of the problem. Alternatively, subjects may insist on the same subjective theorem for the same problem. These hypotheses were also tested in Experiment 1.

Procedure

The subjects in Group 1 were given the original version of the problem (Problem 1) on the first page of the questionnaire, then the modified version (Problem 2) on the second page, and finally the original again on the last

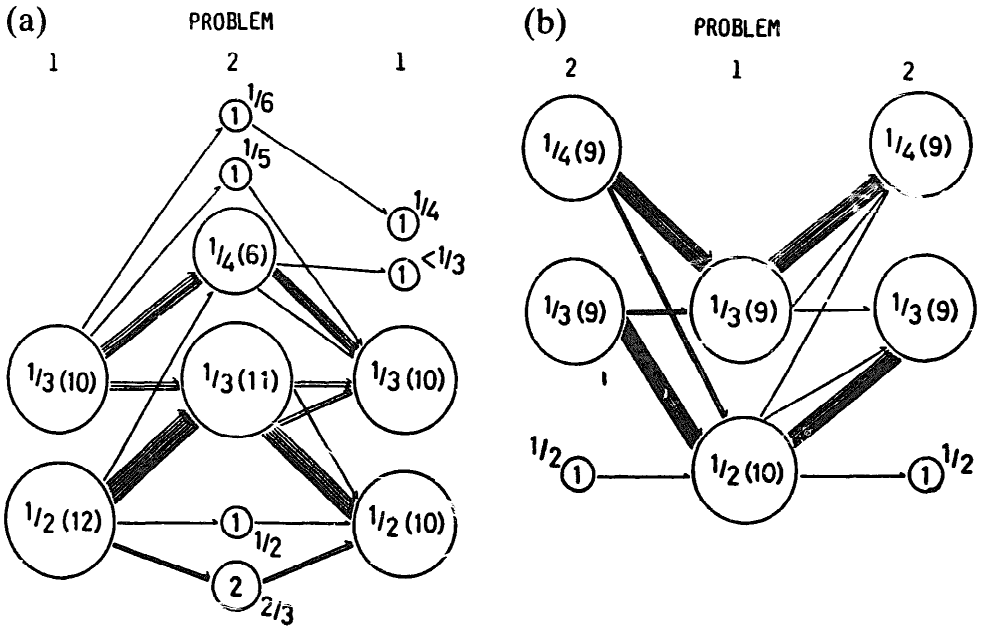
page. Group 2 was given the modified version first, then the original, and finally the modified version again (see Table 2). Subjects were asked to report an intuitive estimate and a brief reason for it to each problem.

Results

Figures 2a and 2b illustrate the results obtained from the subjects in Groups 1 and 2, respectively. The area of each circle represents the number of subjects giving the value as estimate, and arrows indicate the transitions of individual subjects from problem to problem. Additionally, the subjects were categorized into the above-mentioned subjective theorems, based on both their estimates and verbal protocols. The following points are noteworthy:

(1) *Group 1 (Problem 1 → 2 → 1) :* Most subjects (100%, 82% and 91% for each of the three pages) gave probability estimates predicted by three subjective theorems (see Table 1). More than a half of the subjects gave 1/2 as the

Figure 2. *Results of Experiment 1. (a): Group 1. (b): Group 2. The problem number employed on each page of the questionnaire is shown at the top of the figure. The left-most column is for the results on page 1, the middle column for page 2 and the right-most column for page 3 (see Appendix 1 for the problems, Table 2 for the design of experiment). The fraction in each circle represents an estimate for $P(A|b)$. The size of each circle and the number in parenthesis both represent the number of subjects who gave the estimate. The arrows show how the subjects changed their estimates from page to page.*



first intuitive estimate of P(A|b) for the original problem (Problem 1), but it was unclear from most subjects' brief reports whether they used the "number of cases" theorem or the "constant ratio" theorem. However, most of those who gave 1/3, the mathematically "correct" answer, obviously used the "irrelevant, therefore invariant" theorem, judging from their reasoning. Although subjects' estimates of P(A|b) for Problem 2 varied more widely when presented on page 2, 16 out of the 22 subjects repeated their original estimates for the original problem (Problem 1) on page 3. One minor exception was that those who estimated 1/2 again for Problem 1 (on page 3) were more explicit this time in using the "constant ratio" theorem. Interestingly, even when they shifted from one theorem to another for Problem 2 (on page 2), it was not necessarily true that they maintained that second theorem for re-estimating probability for the first problem on page 3.

(2) *Group 2 (Problem 2 → 1 → 2):* When exposed to Problem 2 (the modified version) first, about half of the subjects in Group 2 gave 1/3 as their estimate, as shown in Figure 2b. Unlike the subjects in Group 1, most of them clearly took the "constant ratio" theorem, rather than the "number of cases" theorem from the beginning. The other half of the subjects gave 1/4, taking the "irrelevant, therefore invariant" theorem, and no one gave 1/5, the Bayesian estimate, or anything less than P(A) = 1/4. As with Group 1, the exposure to Problem 1 did not much affect their re-estimation for Problem 2 (the modified version) on page 3. In fact, 17 out of the 19 subjects repeated their original estimates for Problem 2 on page 3.

Discussion

The subjects actually relied on one of the several subjective theorems in estimating P(A|b), judging from the distribution of estimates and their reasoning for them. They seemed to maintain the same theorem, and therefore the same estimate, regardless of their previous exposure to another version of the problem. (As mentioned above, however, in a small number of cases the exposure to Problem 2, i.e., the modified version, drew a subject's attention to the prior probabilities in the problem, and they used the "constant ratio" theorem more explicitly than before for Problem 1, i.e., the original version.) For Problem 1, about a half of the subjects gave the mathematically correct (Bayesian) answer, even though it was for the wrong reason, that is, mostly the "irrelevant, therefore invariant" theorem (note again that the correct Bayesian reasoning does not automatically assume P(A) = P(A|b)). On the other hand, very few subjects gave the mathematically correct answer (1/5) for Problem 2. Thus, the modified version of the three prisoners' problem, in which the prior probabilities were stated explicitly not to be equal, turned out to be even more counterintuitive. The

psychological difficulty with the problems, or the discrepancy between the intuitive and the mathematical estimates of probability, cannot be attributed to the subjects' tendency to neglect the prior probabilities, since most subjects apparently used the prior probabilities to solve Problem 2.

It may still be argued, however, that the subjects relied on the subjective theorems simply because these theorems had been mentioned in the problem itself (in particular, the "irrelevant, therefore invariant," and the "number of cases" theorems were mentioned explicitly). The results of Experiment 2 will show that this interpretation is very unlikely.

Experiment 2

Method

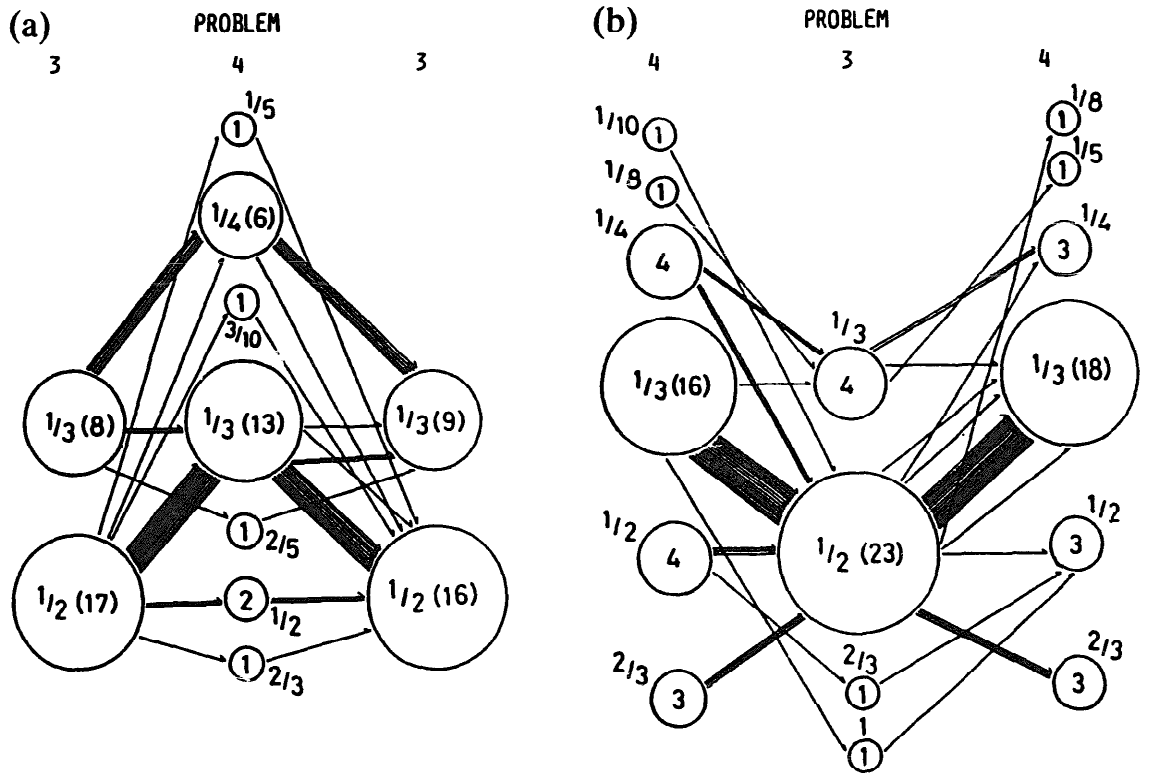
The problems and procedure employed in Experiment 2 were identical to those in Experiment 1 except that all the sentences suggesting possible subjective theorems were eliminated (see Problems 3 and 4 in Appendix 2). The numbers of subjects and the design of the questionnaires are listed in Table 2.

Results and discussion

Figures 3a and 3b show the results for the subjects in Groups 1 and 2. These results are similar to those in Experiment 1 in that the vast majority of responses were like those predicted by one of the subjective theorems. All of the first estimates to Problem 1 given by Group 1 subjects in Experiment 1 (Figure 2a) were those predicted by the theorems, and so were the first estimates to Problem 3 given by Group 1 subjects in Experiment 2 (Figure 3a). Similarly, 100% of the first estimates to Problem 2 given by Group 2 subjects in Experiment 1 (Figure 2b), and 83% of the first estimates to Problem 4 given by Group 2 subjects in Experiment 2 (Figure 3b) were those predicted by the theorems. The difference between the last two cases was not statistically significant ($\chi^2 = 3.66$, $df = 1$). Thus, the subjective theorems were in fact chosen *spontaneously* by the majority of subjects. The discrepancy between the intuitive and Bayesian estimates of probability cannot be attributed to the suggestions which were given in the problem texts used in Experiment 1 (Problems 1 and 2 in Appendix 2).

Two other findings in this experiment were consistent with those in Experiment 1. First, about half of the subjects gave the mathematically correct estimate (1/3) to Problem 3 (the original problem without any suggestions about subjective theorems), but only by taking the "irrelevant, therefore invariant" theorem, judging from their protocols. Very few gave the mathematically correct estimate (1/5) to Problem 4 (the modified version without any suggestions about subjective theorems). The second common

Figure 3. *Results of Experiment 2. (a): Group 1. (b): Group 2. See the caption of Figure 2 for symbols.*



result was that subjects showed a strong tendency to use the same theorem when asked to re-estimate the probability for one version of the problem after being exposed to another version of it. Thus, 22 out of the 25 subjects in Group 1, and 22 out of the 29 subjects repeated their original estimates for Problems 3 and 4, respectively, on page 3.

Experiment 3

The results of these two experiments suggest that the subjective theorems play an important role in intuitive reasoning. Such subjective theorems may be independent of the formal, mathematical process of reasoning by Bayesian theory, but the relationship between the subjective theorems and formal Bayesian reasoning is still unclear. Experiment 3 was designed to examine this issue.

An alternative interpretation of the results of Experiments 1 and 2 is that subjects actually used the Bayesian, or Bayesian-like mathematical scheme of reasoning, but failed to find the correct Bayesian estimate of probability

simply because of some error in applying the formalism (even though the reasoning reported by subjects in Experiments 1 and 2 makes these interpretations unlikely). A more specific version of this interpretation is that subjects could not use the assumption that the jailer has no preference between "B" and "C" to name when both are to be executed (therefore, $P(b|A) = 1/2$; assumption (b) in the Introduction). It may also be that they failed to realize that the jailer has a choice in some cases. If this were so, explicit information on the jailer's choice (an example of what we will call the "Bayesian-critical cue") could markedly reduce subjects' tendency towards the illusory probabilities. Subjects would be very sensitive to such cues and might at least change their estimate of probability even if they still do not give a mathematically correct answer. This was the first prediction to be tested in Experiment 3, and it was done by comparing subjects' estimates for this "explicit, no-preference" version with their estimates for the original, "implicit" version of the problem (Problem 1 in Experiment 1).

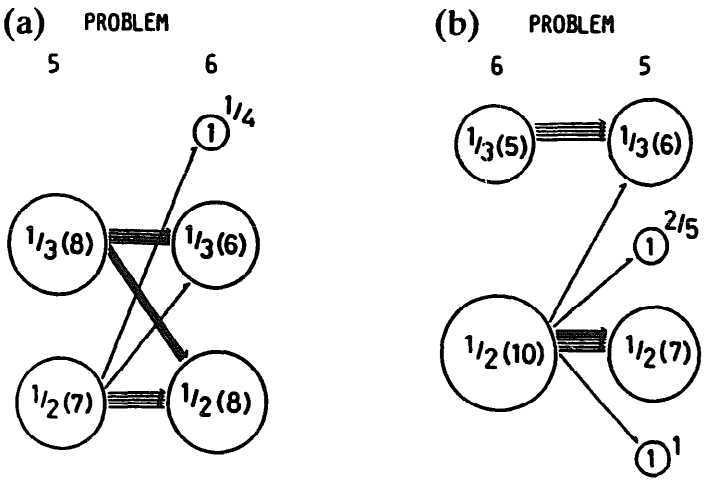
Now, consider an alternative assumption about the jailer's choice: The jailer *always* says "B" when he knows both B and C will be executed. In this case, $P(b|A) = 1$ instead of 1/2 as the assumption (b) stated. The Bayesian estimate of $P(A|b)$ for Mosteller–Lindley's original problem (Problem 1), defined by Equation (3), then turns out to be 1/2. Again, the above-mentioned interpretations predict that subjects would change their estimates of $P(A|b)$ from their estimates with the "explicit, no-preference" assumption. This was the second prediction to be tested, and it was done by comparing subjects' estimates for the "explicit, preference" version with their estimates for the "explicit, no-preference" version of the problem.

Subjects were given one of the two Bayesian-critical probabilities concerning the jailer's choice, and tested for their sensitivity to these probabilities in Experiment 3. If they turn out to be very sensitive to the cues, then the illusory probabilities should be attributed to some error within the Bayesian type of reasoning process. If they are insensitive to the Bayesian-critical cues, then the subjective theorems that subjects have used are more likely to be independent of Bayesian reasoning.

Method

We used the original three prisoners' problem again, but added one of the two notes supplying Bayesian-critical cues at the end of the problem texts (see Problems 5 and 6 in Appendix 2). One note states that the jailer would answer either "B" or "C" with even chances when he has a choice (Problem 5). The other note states that he would always answer "B" when he has a choice between "B" and "C" (Problem 6). Thus, Bayesian estimates of $P(A|b)$ are different in these two problems (1/2 for Problem 5, and 1/3 for

Figure 4. *Results of Experiment 3. (a): Group 1. (b): Group 2. See the caption of Figure 2 for symbols.*



Problem 6). The numbers of subjects and the designs of the questionnaire are shown in Table 2.

Results and discussion

Figures 4a and 4b illustrate the results obtained from the subjects in Groups 1 and 2, respectively. The subjects' estimates of probability again tended to be values which the theorems would predict. Group 1 and Group 2 subjects were pooled together, and classified into three categories: "1/2," "1/3" and "others." A 2 × 3 chi-square test revealed that, against the Bayesian prediction, the distribution of subjects' estimates for Problem 6 was not significantly different from the distribution of first estimates for Problem 1 in Experiment 1 ($\chi^2 = 1.72$, $df = 2$; compare Figures 4a and 4b with Figures 2a and 2b) The subjects were also classified into a 5 × 5 square table by their estimates for Problem 5 and 6 in order to apply a marginal homogeneity test (Everitt, 1977). The results revealed that against the Bayesian prediction, the distribution of subjects' estimates for Problem 6 was not significantly different from the estimate distribution for Problem 5 ($\chi^2 = 4.8$, $df = 4$). Even though 3 out of 15 subjects in Group 1 shifted in the right direction, their reasoning suggested that it had not been because they were sensitive to the Bayesian-critical probabilities (they shifted simply because another theorem came to mind). No noticeable difference was found between the results for Group 1 and Group 2, implying that being exposed to a problem with one Bayesian-critical probability did not much affect the subjects' estimates for a later similar problem with a different Bayesian-critical probability.

These results were less consistent with the hypothesis that the subjects were thinking within a Bayesian-like scheme. The reasons given by the subjects also support this interpretation. Therefore, it is more likely that the inferential schemes or subjective theorems employed by the subjects are different from a Bayesian type of inferential process (see General Discussion, however).

General discussion

Summary of the results and the nature of subjective theorems

(1) *Intuitive difficulty:* The original and the modified version of the problem (Problems 1 and 2 in Appendix 2) turned out to be very difficult for almost all the subjects. For the original problem, about 50% of the subjects in fact gave the mathematically correct estimate of probability, but with "wrong" reasoning. In the modified version of the problem, in which the prior probabilities were explicitly stated, most subjects could not even judge whether the probability would increase or decrease from the prior probability.

(2) *Subjective theorems:* The majority of subjects expressed their intuitive estimation of probability in prototypical schemes of reasoning. These schemes might be based on several distinct subjective theorems, which were not always consistent with the Bayesian framework. Three major subjective theorems were the "number of cases" theorem (case theorem), the "constant ratio" theorem (ratio theorem), and the "irrelevant, therefore invariant" theorem (invariant theorem).

(3) *Spontaneous nature of subjective theorems:* Similar distributions of probability estimates were obtained even when there were no cues to these theorems in the problem text. Thus, the theorems were spontaneously used in intuitive reasoning.

(4) *Persistence of subjective theorems:* Exposure to another variation of the problem did not usually change people's probability estimate and reasoning for a particular problem. Their choice of subjective theorem depended not so much on their experience with a similar problem as on the problem itself.

(5) *Insensitivity to Bayesian-critical cues:* Different additional information as to the jailer's choice, which critically influences the Bayesian estimation, did not significantly affect the subjects' intuitive estimation. Thus, they were insensitive to this kind of Bayesian-critical cue in their intuitive reasoning.

Subjective theorems may be considered as "shorthand" prototypical rules that relate a situation to estimates of probability of particular events. Judging

from our results (1, 2 and 3 above), choosing a subjective theorem and applying it to the problem seem to be spontaneous and common steps in the process of intuitive inference. Our pilot data and Result (4) above strengthen the notion that some cues in the problem context determine which theorem will be chosen from a set of subjective theorems. Based on Result (5) above, it is tempting to conclude that the processes of inferential reasoning with these subjective theorems can be distinctly different from, and in some cases even incompatible with, the Bayesian system of inference. However, it may be also possible that people are in fact capable of using a kind of Bayesian scheme, but only within a simplified framework, as recently suggested by Nisbett and his colleagues (Nisbett, Krantz, Jepson, & Kunda, 1983). For instance, the Bayes' theorem which conditionalizes on B's execution rather than the jailer saying "B will be executed" will not be affected by the jailer's choice,[3] although none of our subjects implied this kind of sample space in their reasoning.

Why are the problems so difficult?

It may be argued that the subjects could not come up with the mathematically correct answer simply because of the time constraints and the instruction to give the intuitive answer. This is very unlikely for two reasons: (a) Subjects (with a reasonable background in statistics) in our pilot study were allowed unlimited time to solve the problem, and yet none of them came up with the Bayesian estimate. (b) Even after the subjects were taught the Bayesian reasoning, most of them reported that it was still counterintuitive. (This was so for subjects in the main experiments, too.) Thus, the amount of required computation for the Bayesian estimates cannot be sufficient to explain the marked discrepancy between the subjects' estimates and the Bayesian estimates.

---

[3]Bayesian estimates conditional on B's execution, rather than the jailer saying "B will be executed," happen to be identical to the estimates given by the ratio theorem, as shown in the following.

For Lindley's original problem (Problem 1), expression (3) (see Introduction) leads to

$$\frac{1\cdot(1/3)}{1\cdot(1/3)+0\cdot(1/3)+1\cdot(1/3)} = 1/2.$$

For the modified version (Problem 2), this same expression is equal to

$$\frac{1\cdot(1/4)}{1\cdot(1/4)+0\cdot(1/4)+1\cdot(1/2)} = 1/3.$$

Since it is not necessary to estimate the conditional probability of the jailer choosing B when A is to be freed, this strategy certainly makes the sample space simpler. Thus, Bayesian computation conditionalized on B's execution may be one example of simplification or "quick aid." (For further discussion, see the subsection "Intuitive resistance to the Bayesian solution" in General Discussion.)

It may be true that since the problem is inherently ambiguous as to the partitioning of a sample space (as exemplified in Footnote 3; also see Nathan, 1986) the subject can hardly come up with the "correct" sample space (Mosteller, 1965). However, the question is what determines the subject's preference for a particular sample space based on a particular subjective theorem.

Tversky and Kahneman (1974, 1982) demonstrated that people have a strong tendency to neglect prior probabilities in their estimation of the posterior probability for Bayesian problems. However, the difficulty with the problem of three prisoners cannot be attributed to this tendency. Both the invariant and ratio theorems, which were used intuitively by most subjects, explicitly utilize the prior probabilities. Further, the Bayesian estimate for the modified version of the problem (Problem 2), which explicitly stated the prior probabilities, was even more counterintuitive than the original version (Problem 1).

The difficulty specific to the "prisoners" problems is caused, at least partially, by the subjects' tendency to ignore *the context* in which the event has occurred. In fact, the Bayesian scheme suggests that the effect of the jailer's answer on the posterior probability varies critically with the way Prisoner A raises the question. For instance, if A's question was simply "Is B to be executed?" and the jailer's answer was "Yes" in the modified version of the problem (Problem 2), the posterior probability of A's survival would be:

$$P(A|b) = \frac{1\cdot(1/4)}{1\cdot(1/4) + 0\cdot(1/4) + 1\cdot(1/2)} = 1/3,$$

which is equal to the estimate given by the ratio theorem (see Table 1). Also, when A asks the jailer to randomly name a man to be executed among the three prisoners (including A himself) and the jailer answers "B is," the posterior probability would be:

$$P(A|b) = \frac{(1/2)\cdot(1/4)}{(1/2)\cdot(1/4) + 0\cdot(1/4) + (1/2)\cdot(1/2)} = 1/3.$$

Again, this is the same as the estimate given by the ratio theorem. Note that in both cases the statement given by the jailer to A is identical, and yet the Bayesian estimate varies depending on the context in which the statement is given.

Based on these theoretical observations, we suspect, although merely as a conjecture, that a considerable proportion of the subjects insisted on the ratio theorem and the estimate given by it without realizing that this scheme is *not always* appropriate because: (1) they neglected the critical relevance of the context of the question, (2) the ratio theorem functioned well, without re-

garding the context of the question, as a rule of thumb to make an instant estimate of probability, and (3) the estimate based on the ratio theorem agrees well with the realistic or appropriate estimate (i.e., the Bayesian estimate) *unless* a specific question is asked, as in Problem 1 and 2 in the current study. This notion is further supported by the fact that most subjects insisted on the subjective theorems, none of which actually take into account the context in which the data were given.

The estimates based on the ratio theorem match the Bayesian probabilities when $P(b|A) = P(b|C)$, as obvious from the above-mentioned variations of the problem. One may suspect that the difficulty of the prisoner problems is predominantly due to the failure to be aware of the values of $P(b|A)$ and $P(b|C)$. When A is the man to be set free, the jailer can choose to name either B or C, thus $P(b|A) = 1/2$; whereas when C is to be set free, then B is the only man of B and C who is to be executed, leaving the jailer with no other choice than B to name. Thus, there is an alternative explanation for the biased estimation: Subjects were thinking within the Bayesian, or Bayesian-like scheme, but simply did not realize that the jailer's choice is critically relevant. Our Result (5), however, suggests that this is unlikely to be the cause of illusory probabilities. Our subjects were not sensitive to this type of Bayesian-critical cue in Experiment 3. Additional explicit information about the relationship between $P(b|A)$ and $P(c|A)$ did not change their intuitive inference. Subjects gave the "wrong" estimates and reasoning not because they did not realize the inequality of $P(b|A)$ and $P(b|C)$, but rather, possibly because they did not realize that a particular event affects the posterior probability in different ways, depending on the situational context. That is, they could not realize that a conditional event is defined not merely by the information obtained, but also by the way in which it was obtained (Bar-Hillel & Falk, 1982). This context-dependency may be better understood in a Bayesian reasoning scheme, in which the influence of a particular event is specifically related to what else could have happened and the probabilities of those alternative events.

The strong tendency to neglect this context-dependency may be partially caused by subjects' knowledge that the value (true/false) of a proposition ("B is to be executed") cannot be altered by the question to which the proposition is a response. (Compare the modified version employed in the experiments and the two examples mentioned above, again. Regardless of A's question, the proposition that B is to be executed is always true. See also Footnote 3.)

Intuitive resistance to the Bayesian solution

We have so far discussed what prevents people from utilizing the Bayesian reasoning scheme. There are two important unresolved issues: (1) Why do people use just a few subjective theorems more or less unanimously? (2) Why is the Bayesian solution so counterintuitive even after we logically understand it?

There seems to be a hierarchical structure of formulated beliefs (theorems) underlying the subjects' probability reasoning, which is similar to what Shafer & Tversky (1985) found in their study of "languages of belief functions" for probability judgment. For example, some subjective theorems may be supported by reasoning such as: "Now that one of Prisoner A's two rivals has been kicked out, A's chances of survival can by no means decrease." Note that this reasoning is consistent with *all* of the subjective theorems, but *not* with the Bayesian solution. The kind of belief underlying this reasoning may be called a "superior" subjective theorem in the sense that it functions as a crude, but stronger rule which tests, selects and supports other subjective theorems. For another example, the ratio and case theorems support each other in the original problem because they give the same estimate of probability (see Discussion of Experiment 1), but inhibit each other in the modified version (Problem 2). Formulation of such structure among beliefs has been suggested not only in human reasoning processes, but also in artificial intelligence expert systems (e.g., Duda, Hart, & Nilsson, 1976).

Subjective theorems are simplifications which aid us in quickly finding mathematically correct answers for various versions of the original problem. Appendix 1 explicitly shows general conditions under which $P(A|b)$ is unchanged from $P(A)$. Thus, when $P(B) = P(C)$, the invariant theorem works as a quick method to estimate $P(A|b)$.

More generally, the three prisoners' problem consists of three elements of information: (1) the prior probabilities, (2) prisoner A's question and (3) the jailer's answer. (The second element cannot be neglected because of the context-dependency mentioned in the previous section.) Degenerate versions of the problem can be systematically created by omitting one or two of these elements. In fact, the original problem is a degenerate version of our modified version because the original is created by omitting the prior probabilities from the modified version. The invariant theorem coincides with the Bayesian estimate of $P(A|b) = 1/3$. Alternatively, if A does not ask the question, but the jailer simply tells A that B is going to be executed, then $P(b|A) = P(b|C)$ and $P(b|B) = 0$. The Bayesian estimate would be 1/3 in the modified version, coincident with the ratio theorem, as mentioned in the previous subsection. As indicated in these examples, the theorems can be good bases for rules of thumb for many naturally occurring problems. Since subjective theorems are

functionally "valid" in most cases in real life, they are constantly reinforced, and thus, survive. From this viewpoint, the "mistakes" that our subjects have made can be regarded as a consequence of overgeneralization or misapplication of these rules. Even though this overgeneralization is a general cause of cognitive fallacy, which is not specific at all to Bayesian probabilities, the Bayesian problems may be psychologically the most intriguing examples because fully understanding the Bayesian reasoning sometimes won't totally exorcise the "illusory" judgment (see the subsection "Pilot study" in Introduction).

Finally, there is some general difficulty in learning any kind of probabilistic reasoning scheme. Since it is a matter of probability, feedback from several outcomes would not be sufficient to reinforce or punish a particular strategy of decision making. As an example, the experience of A being set free twice out of five identical occasions would not indicate whether 1/5, 1/3, or even 1/2 is the mathematically correct answer. Also, it should be noted that even though it is necessary to collect and compare identical situations to learn the appropriate estimate of probability purely empirically, it is practically impossible to prove that similar situations are in fact identical.

Despite these difficulties, some general cognitive schemes are useful for integrating the Bayesian scheme into our intuitive system of reasoning. For instance, it may be helpful to consider the theorem of total probability, which states that the probability of an event is considered to be a weighted average of posterior probabilities of exhaustive and exclusive cases. According to this theorem,

$$\begin{aligned} P(A) &= P(A|a)P(a) + P(A|b)P(b) + P(A|c)P(c) \\ &= P(A|b)P(b) + P(A|c)P(c). \end{aligned}$$

This equation indicates that $P(A)$ is a weighted average of $P(A|b)$ and $P(A|c)$ since $P(b) + P(c) = 1$. Thus, only three cases are conceivable:

$$\begin{aligned} &P(A) > P(A|b), \text{ and } P(A) < P(A|c) \\ &P(A) < P(A|b), \text{ and } P(A) > P(A|c) \\ &P(A) = P(A|b) = P(A|c). \end{aligned}$$

As obvious from this, A's chance to be set free can *not always* increase *regardless* of whether the jailer names B or C to be executed. This may possibly help to reject the above-mentioned "superior" theorem that when an alternative is eliminated, the probabilities of the other alternatives by no means decrease.

Appendix 1. The condition for $P(A|b) \leq P(A)$

As detailed in the text, $P(A|b)$ in the three prisoners' problem can be expressed as

$$P(A|b) = \frac{P(b|A)P(A)}{P(b|A)P(A) + P(b|B)P(B) + P(b|C)P(C)} \tag{3}$$

where we could assume

$$P(b|A) = 1/2, \quad P(b|B) = 0, \quad P(b|C) = 1. \tag{2}$$

Inserting Equation (2) into (3), we obtain

$$P(A|b) = \frac{P(A)}{P(A) + 2\{P(C)\}}. \tag{6}$$

By the way, it was given in the problem that

$$P(A) + P(B) + P(C) = 1. \tag{7}$$

From Equation (6) and (7), we can obtain the new expression of $P(A|b)$:

$$P(A|b) = \frac{P(A)}{\{1 - P(B) - P(C)\} + 2\{P(C)\}} = \frac{P(A)}{1 - P(B) + P(C)}. \tag{8}$$

It is now obvious from Equation (8) that $P(A|b) \leq P(A)$ when $P(B) \leq P(C)$. On the contrary, $P(A|b) > P(A)$ when $P(B) > P(C)$. Thus, whenever the jailer guarantees that the "weaker" of the two rivals will be executed, A's chance will decrease. Interestingly, it is totally irrelevant whether P(B) or P(C), the prior probabilities of B or C, is greater than P(A). It is just a matter of comparison between P(B) and P(C).

Appendix 2. Problems and questions

Problem 1: (See the Introduction for the problem.)
Questions: (1) How would you estimate the probability of A to be freed after the jailer's answer? (2) Show the reason briefly.

Problem 2: (See the Introduction for the problem.)
Questions: (The same as those for Problem 1.)

Problem 3: Three men, A, B and C were in jail. A knew that one of them was to be set free and the other two were to be executed. But he didn't know who was the one to be spared. To the jailer who did know, A asked "Will

you tell me the name of one man, B or C, who is going to be executed?" The jailer answered, "B will be executed."
Questions: (The same as those for Problem 1.)

Problem 4: Three men, A, B and C were in jail. One of them was to be set free and the other two were to be executed. A had reason to believe that their chances of being freed were: A:1/4, B:1/4, C:1/2. After their fates had been decided, A who didn't know the outcome of decision asked the jailer who did, "Will you tell me the name of one man, B or C, who is going to be executed?" The jailer answered, "B will be executed."
Questions: (The same as those for Problem 1.)

Problem 5: (Exactly the same as Problem 1 except the following note which was added at the end.) *Note:* The jailer would answer either "B" or "C" with even chances when he has a choice between "B" and "C."
Questions: (The same as Problem 1.)

Problem 6: (Exactly the same as Problem 1 except the following note which was added at the end.) *Note:* The jailer would always answer "B" when he has a choice between "B" and "C."
Questions: (The same as Problem 1.)

References

Bar-Hillel, M., & Falk, R. (1982). Some teasers concerning conditional probabilities. *Cognition, 11,* 109–122.
Duda, R.O., Hart, P.E., & Nilsson, N.J. (1976). *Subjective Bayesian methods for rule-based inference systems* (Technical Note 124). Menlo Park, CA: Stanford Research Institute.
Everitt, B.S. (1977). *The analysis of contingency tables.* London: Chapman and Hall.
Lindley, D.V. (1971). *Making decisions.* London: John Wiley.
Mosteller, F. (1965). *Fifty challenging problems in probability with solutions.* Reading, MA: Addison-Wesley.
Nathan, A. (1986). How not to solve it. *Philosophy of Science, 53,* 114–116.
Nisbett, R.E., Krantz, D.H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review, 90,* 339–363.
Shafer, G., & Tversky, A. (1985). Languages and designs for probability judgment. *Cognitive Science, 9,* 309–339.
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131.
Tversky, A., & Kahneman, D. (1982). Evidential impacts of base rates. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.

Résumé

Parmi les divers problèmes probabilités Bayesiens, le "problème des trois prisonniers" (Mosteller, 1965; Lindley, 1971) est un exemple qui illustre particulièrement bien l'incompatibilité de deux formes de raisonnement dans le domaine des probabilités: le raisonnement intuitif d'une part et le raisonnement mathématique forme d'autre part. Le problème, en particulier, éveille notre curiosité à l'égard des questions concernant l'influence que peuvent avoir dans les domaines mathématique et cognitif de facteurs tels que les "probabilités à priori" et le contexte dans lequel est donnée une certaine information. Dans le présent article, on expose une nouvelle version du problème dont la solution contredit notre intuition de façon encore plus marquée. Cette nouvelle version fut élaborée également afin que différents schémas inférentiels aboutissent à des estimations de "probabilités à posteriori" distinctes les unes des autres. Les données obtenues à partir de questionnaires et d'analyses théoriques des versions originales et modifiées du problème suggèrent que: (1) Les processus psychologiques du raisonnement intuitif sont qualitativement différents de ceux du raisonnement mathématique. (2) La tendance à négliger les "probabilités à priori" (Kahneman & Tversky 1974, 1982) ne constitute pas toujours le facteur déterminant pour ce qui est des erreurs de jugement. (3) On peut faire une partition des jugements intuitifs selon la "croyance propositionnelle" dont ils dérivent. Ces diverses croyances au caractère inconsistant et grossier seront appelées "théorèmes subjectifs"; on discutera dans le présent article de leur nature et de leurs rôles.