PROBABLE KNOWLEDGE

RICHARD C. JEFFREY
The City University of New York

The central problem of epistemology is often taken to be that of explaining how we can know what we do, but the content of this problem changes from age to age with the scope of what we take ourselves to know; and philosophers who are impressed with this flux sometimes set themselves the problem of explaining how we can get along, knowing as little as we do. For knowledge is sure, and there seems to be little we can be sure of outside logic and mathematics and truths related immediately to experience. It is as if there were some propositions – that this paper is white, that two and two are four – on which we have a firm grip, while the rest, including most of the theses of science, are slippery or insubstantial or somehow inaccessible to us. Outside the realm of what we are sure of lies the puzzling region of probable knowledge – puzzling in part because the sense of the noun seems to be cancelled by that of the adjective.

The obvious move is to deny that the notion of knowledge has the importance generally attributed to it, and to try to make the concept of belief do the work that philosophers have generally assigned the grander concept. I shall argue that this is the right move.

1. *A pragmatic analysis of belief.* To begin, we must get clear about the relevant sense of 'belief'. Here I follow Ramsey: 'the kind of measurement of belief with which probability is concerned is ... a measurement of belief *qua* basis of action' [1].

[1] Frank P. Ramsey, 'Truth and probability', in *The Foundations of Mathematics and Other Logical Essays*, R. B. Braithwaite, ed., London and New York, 1931, p. 171.

Ramsey's basic idea was that the desirability of a gamble G is a weighted average of the desirabilities of winning and of losing in which the weights are the probabilities of winning and of losing. If the proposition gambled upon is A, if the prize for winning is the truth of a proposition W, and if the penalty for losing is the truth of a proposition L, we then have

$$(1) \qquad prob\ A = \frac{des\ G - des\ L}{des\ W - des\ L}.$$

Thus, if the desirabilities of losing and of winning happen to be 0 and 1, we have $prob\ A = des\ G$, as illustrated in Figure 1, for the case in which the probability of winning is thought to be $\frac{3}{4}$.

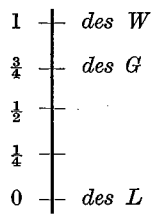

Figure 1.

On this basis, Ramsey [1] is able to give rules for deriving the gambler's subjective probability *and* desirability functions from his preference ranking of gambles, provided the preference ranking satisfies certain conditions of consistency. The probability function obtained in this way is a probability measure in the technical sense that, given any finite set of pairwise incompatible propositions which together exhaust all possibilities, their probabilities are non-negative real numbers that add up to 1. And in an obvious sense, probability so construed is a measure of the subject's willingness to act on his beliefs in propositions: it is a measure of degree of belief.

I propose to use what I take to be an improvement of Ramsey's scheme, in which the work that Ramsey does with the operation

[1] 'Truth and probability', F. P. Ramsey, *op. cit.*

of forming gambles is done with the usual truth-functional operations on propositions [1]. The basic move is to restrict attention to certain 'natural' gambles, in which the prize for winning is the truth of the proposition gambled upon, and the penalty for losing is the falsity of that proposition. In general, the situation in which the gambler takes himself to be gambling on A with prize W and loss L is one in which he believes the proposition

$$G = AW \vee \bar{A}L.$$

If G is a natural gamble we have $W = A$ and $L = \bar{A}$, so that G is the necessary proposition, $T = A \vee \bar{A}$:

$$G = AA \vee \bar{A}\bar{A} = T.$$

Now if A is a proposition which the subject thinks good (or bad) in the sense that he places it above T (or below T) in his preference ranking, we have

$$prob\ A = \frac{des\ T - des\ \bar{A}}{des\ A - des\ \bar{A}}, \tag{2}$$

corresponding to Ramsey's formula (1).

Here the basic idea is that if $A_1, A_2, \ldots, A_n$ are an exhaustive set of incompatible ways in which the proposition A can come true, the desirability of A must be a weighted average of the desirabilities of the ways in which it can come true:

$$des\ A = w_1\ des\ A_1 + w_2\ des\ A_2 + \ldots + w_n\ des\ A_n, \tag{3}$$

where the weights are the conditional probabilities,

$$w_i = prob\ A_i / prob\ A. \tag{4}$$

Let us call a function *des* which attributes real numbers to propositions a *Bayesian desirability function* if there is a probability measure *prob* relative to which (3) holds for all suitable $A, A_1, A_2, \ldots, A_n$. And let us call a preference ranking of propositions *coherent* if there is a Bayesian desirability function which ranks those propositions in order of magnitude exactly as they are ranked in order of preference. One can show [1] that if certain weak conditions are met by a coherent preference ranking, the underlying desirability function is determined up to a fractional linear transformation, i.e., if *des* and *DES* both rank propositions in order of magnitude exactly as they are ranked in order of preference, there must be real numbers a, b, c, d such that for any proposition A in the ranking we have

$$DES\ A = \frac{a\ des\ A + b}{c\ des\ A + d}. \tag{5}$$

The probability measure *prob* is then determined by (2) up to a certain quantization. In particular, if *des* is Bayesian relative to *prob*, then *DES* will be Bayesian relative to *PROB*, where

$$PROB\ A = prob\ A\ (c\ des\ A + d). \tag{6}$$

Under further plausible conditions, (5) and (6) are given either exactly (as in Ramsey's theory) or approximately by

$$DES\ A = a\ des\ A + b, \tag{7}$$

$$PROB\ A = prob\ A. \tag{8}$$

I take the principal advantage of the present theory over Ramsey's to be that here we work with the subject's actual beliefs, whereas Ramsey needs to know what the subject's preference ranking of relevant propositions would be if his views of what the world is were to be changed by virtue of his having come to believe that various arbitrary and sometimes bizarre causal relationships had been established via gambles [2].

To see more directly how preferences may reflect beliefs in the present system, observe that by (2) we must have

[1] See Richard C. Jeffrey, *The Logic of Decision*, McGraw-Hill, 1965, the mathematical basis for which can be found in Ethan Bolker, *Functions Resembling Quotients of Measures*, Ph. D. Dissertation, Harvard University, 1965, and *Trans. Am. Math. Soc.*, 124, 1966, pp. 293–312.

[1] Jeffrey, *op. cit.*, chs. 6, 8.

[2] Jeffrey, *op. cit.*, pp. 145–150.

prob $A > prob\ B$ if the relevant portion of the preference ranking is

$$\begin{matrix} A, & B \\ T & \\ \bar{B} & \\ \bar{A} & \end{matrix}$$

In particular, suppose that A and B are the propositions that the subject will get job 1 and that he will get job 2, respectively. Pay, working conditions, etc., are the same, so that he ranks A and B together. Now if he thinks himself more likely to get job 1 than job 2, he will prefer a guarantee of $(\bar{B})$ not getting job 2 to a guarantee of $(\bar{A})$ not getting job 1; for he thinks that an assurance of not getting job 2 leaves him more likely to get one or the other of the equally liked jobs than would an assurance of not getting job 1.

2. *Probabilistic acts and observations.* We might call a proposition *observational* for a certain person at a certain time if at that time he can make an observation of which the *direct* effect will be that his degree of belief in the proposition will change to 0 or to 1. Similarly, we might call a proposition *actual* for a certain person at a certain time if at that time he can perform an act of which the *direct* effect will be that his degree of belief in the proposition will change to 0 or to 1. Under ordinary circumstances, the proposition that the sun is shining is observational and the proposition that the agent blows his nose is actual. Performance of an act may give the agent what Anscombe calls [1] 'knowledge without observation' of the truth of an appropriate actual proposition. Apparently, a proposition can be actual or observational without the agent's knowing that it is; and the agent can be mistaken in thinking a proposition actual or observational.

The point and meaning of the requirement that the effect be 'direct', in the definitions of 'actual' and 'observational', can be illustrated by considering the case of a sleeper who awakens and

[1] G. E. M. Anscombe, *Intention*, § 8, Oxford, 1957; 2nd ed., Ithaca, N.Y., 1963.

sees that the sun is shining. Then one might take the observation to have shown him, directly, that the sun is shining, and to have shown him indirectly that it is daytime. In general, an observation will cause numerous changes in the observer's belief function, but many of these can be construed as consequences of others. If there is a proposition E such that the *direct* effect of the observation is to change the observer's degree of belief in E to 1, then for any proposition A in the observer's preference ranking, his degree of belief in A after the observation will be the conditional probability

$$prob_E\ A = prob\ (A/E) = prob\ AE/prob\ E, \tag{9}$$

where *prob* is the observer's belief function before the observation. And conversely, if the observer's belief function after the observation is $prob_E$ and $prob_E$ is not identical with *prob*, then the *direct* effect of the observation will be to change the observer's degree of belief in E to 1. This completes a definition of *direct*.

But from a certain strict point of view, it is rarely or never that there is a proposition for which the direct effect of an observation is to change the observer's degree of belief in that proposition to 1; and from that point of view, the classes of propositions that count as observational or actual in the senses defined above are either empty or as good as empty for practical purposes. For if we care seriously to distinguish between 0.999 999 and 1.000 000 as degrees of belief, we may find that, after looking out the window, the observer's degree of belief in the proposition that the sun is shining is not quite 1, perhaps because he thinks there is one chance in a million that he is deluded or deceived in some way; and similarly for acts where we can generally take ourselves to be at best *trying* (perhaps with very high probability of success) to make a certain proposition true.

One way in which philosophers have tried to resolve this difficulty is to postulate a phenomenalistic language in which an appropriate proposition E can always be expressed, as a report on the immediate content of experience; but for excellent reasons, this move is now in low repute [1]. The crucial point is not that 0.999 999 is so close

[1] See, e.g., J. L. Austin, *Sense and Sensibilia*, Oxford, 1962.

to 1.000 000 as to make no odds, practically speaking, for situations abound in which the gap is more like one half than one millionth. Thus, in examining a piece of cloth by candlelight one might come to attribute probabilities 0.6 and 0.4 to the propositions G that the cloth is green and B that it is blue, without there being any proposition E for which the direct effect of the observation is anything near changing the observer's degree of belief in E to 1. One might think of some such proposition as that (E) *the cloth looks green or possibly blue*, but this is far too vague to yield *prob* $(G/E)=0.6$ and *prob* $(B/E)=0.4$. Certainly, there is *something* about what the observer sees that leads him to have the indicated degrees of belief in G and in B, but there is no reason to think the observer can express this something by a statement in his language. And physicalistically, there is some perfectly definite pattern of stimulation of the rods and cones of the observer's retina which prompts his belief, but there is no reason to expect him to be able to describe that pattern or to recognize a true description of it, should it be suggested.

As Austin [1] points out, the crucial mistake is to speak seriously of the *evidence* of the senses. Indeed the relevant experiences have perfectly definite characteristics by virtue of which the observer comes to believe as he does, and by virtue of which in our example he comes to have degree of belief 0.6 in G. But it does not follow that there is a proposition E of which the observer is certain after the observation and for which we have *prob* $(G/E)=0.6$, *prob* $(B/E)=0.4$, etc.

In part, the quest for such phenomenological certainty seems to have been prompted by an inability to see how uncertain evidence can be used. Thus C. I. Lewis:

> If anything is to be probable, then something must be certain. The data which themselves support a genuine probability, must themselves be certainties. We do have such absolute certainties, in the sense data initiating belief and in those passages of experience which later may confirm it. But neither such initial data nor such later verifying passages of experience can be phrased in the language of objective statement – because what can be so phrased is never more than probable. Our sense certainties can only be formulated by the expressive use of language, in which what is signified is a content of experience and what is asserted is the givenness of this content [1].

[1] Austin, *op. cit.*, ch. 10.

But this motive for the quest is easily disposed of [2]. Thus, in the example of observation by candlelight, we may take the direct result of the observation (in a modified sense of 'direct') to be, that the observer's degrees of belief in G and B change to 0.6 and 0.4. Then his degree of belief in any proposition A in his preference ranking will change from *prob* A to

$$PROB\ A=0.6\ prob\ (A/G)+0.4\ prob\ (A/B).$$

In general, suppose that there are propositions $E_1, E_2, \ldots, E_n$, in which the observer's degrees of belief after the observation are $p_1, p_2, \ldots, p_n$; where the E's are pairwise incompatible and collectively exhaustive; where for each i, *prob* E_i is neither 0 nor 1; and where for each proposition A in the preference ranking and for each i the conditional probability of A on E_i is unaffected by the observation:

$$PROB\ (A/E_i)=prob\ (A/E_i). \tag{10}$$

Then the belief function after the observation may be taken to be *PROB*, where

$$PROB\ A=p_1\ prob\ (A/E_1)+p_2\ prob\ (A/E_2)+\ldots +p_n\ prob\ (A/E_n), \tag{11}$$

if the observer's preference rankings before and after the observation are both coherent. Where these conditions are met, the propositions $E_1, E_2, \ldots, E_n$, may be said to form a *basis* for the observation; and the notion of a basis will play the role vacated by the notion of *directness*.

[1] C. I. Lewis, *An Analysis of Knowledge and Valuation*, La Salle, Illinois, 1946, p. 186.

[2] Jeffrey, *op. cit.*, ch. 11.

The situation is similar in the case of acts. A marksman may have a fairly definite idea of his chances of hitting a distant target, e.g. he may have degree of belief 0.3 in the proposition H that he will hit it. The basis for this belief may be his impressions of wind conditions, quality of the rifle, etc.; but there need be no reason to suppose that the marksman can express the relevant data; nor need there be any proposition E in his preference ranking in which the marksman's degree of belief changes to 1 upon deciding to fire at the target, and for which we have *prob* $(H/E)=0.3$. But the pair H, $\bar{H}$ may constitute a *basis* for the act, in the sense that for any proposition A in the marksman's preference ranking, his degree of belief after his decision is

$$PROB\ A = 0.3\ prob\ (A/H) + 0.7\ prob\ (A/\bar{H}).$$

It is correct to describe the marksman as *trying* to hit the target; but the proposition that he is trying to hit the target can not play the role of E above. Similarly, it was correct to describe the cloth as *looking* green or possibly blue; but the proposition that the cloth looks green or possibly blue does not satisfy the conditions for directness.

The notion of directness is useful as well for the resolution of unphilosophical posers about probabilities, in which the puzzling element sometimes consists in failure to think of an appropriate proposition E such that the direct effect of an observation is to change degree of belief in E to 1, e.g. in the following problem reported by Mosteller [1].

> Three prisoners, a, b, and c, with apparently equally good records have applied for parole. The parole board has decided to release two of the three, and the prisoners know this but not which two. A warder friend of prisoner a knows who are to be released. Prisoner a realizes that it would be unethical to ask the warder if he, a, is to be released, but thinks of asking for the name of *one* prisoner *other than himself* who is to be released. He thinks that before he asks, his chances of release are $\frac{2}{3}$. He thinks that if the warder says 'b will be released,' his own chances have now gone down to $\frac{1}{2}$, because either a and b or b and c are to be released. And so a decides not to reduce his chances by asking. However, a is mistaken in his calculations. Explain.

[1] Problem 13 of Frederick Mosteller, *Fifty Challenging Problems in Probability*, Reading, Mass., Palo Alto, and London, 1965.

Here indeed the possible cases (in a self-explanatory notation) are

$$AB,\ AC,\ BC,$$

and these are viewed by a as equiprobable. Then *prob* A is $\frac{2}{3}$ but *prob* $(A/B)=$ *prob* $(A/C)=\frac{1}{2}$, and, since the warder must answer either 'b' or 'c' to a's question, it looks as if the direct result of the 'observation' will be that a comes to attribute probability 1 either to the proposition B that b will be released, or to the proposition C that c will be released. But this is incorrect. The relevant evidence-proposition would be more like the proposition *that the warder says*, 'b', or *that the warder says*, 'c', even though neither of these will quite do. For it is only in cases AB and AC that the warder's reply is dictated by the facts: in case BC, where b and c are both to be released, the warder must somehow choose *one* of the two true answers. If a expects the warder to make the choice by some such random device as tossing a coin, then we have *prob* (A/the warder says, 'b') = *prob* (A/the warder says, 'c') = = *prob* $A=\frac{2}{3}$; while if a is sure that the warder will say 'b' if he can, we have *prob* (A/the warder says 'b') $=\frac{1}{2}$ but *prob* (A/the warder says 'c') $=1$.

3. *Belief: reasons vs. causes.* Indeed it is desirable, where possible, to incorporate the results of observation into the structure of one's beliefs via a basis of form E, $\bar{E}$ where the probability of E after the observation is nearly 1. For practical purposes, E then satisfies the conditions of directness, and the 'direct' effect of the observation can be described as informing the observer of the truth of E. Where this is possible, the relevant passage of sense experience *causes* the observer to believe E; and if *prob* (A/E) is high, his belief in E may be a *reason* for his believing A, and E

may be spoken of as (inconclusive) *evidence* for A. But the sense experience is evidence neither for E nor for A. Nor does the situation change when we speak physicalistically in terms of patterns of irritation of our sensory surfaces, instead of in terms of sense experience: such patterns of irritation *cause* us to believe various propositions to various degrees; and sometimes the situation can be helpfully analyzed into one in which we are caused to believe $E_1, E_2, \ldots, E_n$, to degrees $p_1, p_2, \ldots, p_n$, whereupon those beliefs provide *reasons* for believing other propositions to other degrees. But patterns of irritation of our sensory surfaces are not reasons or evidence for any of our beliefs, any more than irritation of the mucous membrane of the nose is a *reason* for sneezing.

When I stand blinking in bright sunlight, I can no more believe that the hour is midnight than I can fly. My degree of belief in the proposition that the sun is shining has two distinct characteristics. (a) It is 1, as close as makes no odds. (b) It is compulsory. Here I want to emphasize the second characteristic, which is most often found in conjunction with the first, but not always. Thus, if I examine a normal coin at great length, and experiment with it at length, my degree of belief in the proposition that the next toss will yield a head will have two characteristics. (a) It is $\frac{1}{2}$. (b) It is compulsory. In the case of the coin as in the case of the sun, I cannot decide to have a different degree of belief in the proposition, any more than I can decide to walk on air.

In my scientific and practical undertakings I must make use of such compulsory beliefs. In attempting to understand or to affect the world, I cannot escape the fact that I am part of it: I must rather make use of that fact as best I can. Now where epistemologists have spoken of observation as a source of *knowledge*, I want to speak of observation as a source of compulsory *belief* to one or another degree. I do not propose to identify a very high degree of belief with knowledge, any more than I propose to identify the property of being near 1 with the property of being compulsory.

Nor do I postulate any *general* positive or negative connection between the characteristic of being compulsory and the characteristic of being sound or appropriate in the light of the believer's experience. Nor, finally, do I take a compulsory belief to be necessarily a permanent one: new experience or new reflection (perhaps, prompted by the arguments of others) may loosen the bonds of compulsion, and may then establish new bonds; and the effect may be that the new state of belief is sounder than the old, or less sound.

Then why should we trust our beliefs? According to K. R. Popper,

> . . . the decision to accept a basic statement, and to be satisfied with it, is causally connected with our experiences – especially with our *perceptual experiences*. But we do not attempt to *justify* basic statements by these experiences. Experiences can *motivate a decision*, and hence an acceptance or a rejection of a statement, but a basic statement cannot be *justified* by them – no more than by thumping the table [1].

I take this objection to be defective, principally in attempting to deal with basic statements (observation reports) in terms of *decisions* to *accept* or to *reject* them. Here acceptance parallels belief, rejection parallels disbelief (belief in the denial), and tentativeness or reversibility of the decision parallels *degree* of belief. Because logical relations hold between statements, but not between events and statements, the relationship between a perceptual experience (an event of a certain sort) and a basic statement cannot be a logical one, and therefore, Popper believes, cannot be of a sort that would justify the statement:

> Basic statements are accepted as the result of a decision or agreement; and to that extent they are conventions [2].

But in the absence of a positive account of the nature of acceptance and rejection, parallel to the account of partial belief given in section 1, it is impossible to evaluate this view. Acceptance and rejection are apparently acts undertaken as results of decisions; but somehow the decisions are conventional – perhaps only in the sense that they may be *motivated* by experience, but not *adequately* motivated, if adequacy entails justification.

---

[1] K. R. Popper, *The Logic of Scientific Discovery*, London, 1959, p. 105.
[2] Popper, *op. cit.*, p. 106.

To return to the question, 'Why should we trust our beliefs?' one must ask what would be involved in *not* trusting one's beliefs, if belief is analyzed as in section **1** in terms of one's preference structure. One way of mistrusting a belief is declining to act on it, but this appears to consist merely in lowering the degree of that belief: to mistrust a partial belief is then to alter its degree to a new, more suitable value.

A more hopeful analysis of such mistrust might introduce the notion of sensitivity to further evidence or experience. Thus, agents 1 and 2 might have the same degree of belief – $\frac{1}{2}$ – in the proposition H_1 that the first toss of a certain coin will yield a head, but agent 1 might have this degree of belief because he is convinced that the coin is normal, while agent 2 is convinced that it is either two-headed or two-tailed, he knows not which [1]. There is no question here of agent 2's expressing his mistrust of the figure $\frac{1}{2}$ by lowering or raising it, but he can express that mistrust quite handily by aspects of his belief function. Thus, if H_i is the proposition that the coin lands head up the ith time it is tossed, agent 2's beliefs about the coin are accurately expressed by the function $prob_2$ where

$$prob_2\ H_i = \tfrac{1}{2},\ prob_2\ (H_i/H_j) = 1,$$

while agent 1's beliefs are equally accurately expressed by the function $prob_1$ where

$$prob_1\ (H_{i_1}, H_{i_2}, \ldots, H_{i_n}) = 2^{-n},$$

if $i_1 < i_2 < \ldots < i_n$. In an obvious sense, agent 1's beliefs are *firm* in the sense that he will not change them in the light of further evidence, since we have

$$prob_1\ (H_{n+1}/H_1, H_2, \ldots, H_n) = prob_1\ H_{n+1} = \tfrac{1}{2},$$

while agent 2's beliefs are quite tentative and in that sense, mistrusted by their holder. Still, $prob_1\ H_i = prob_2\ H_i = \frac{1}{2}$.

After these defensive remarks, let me say how and why I take

[1] This is a simplified version of 'the paradox of ideal evidence', Popper, *op. cit.*, pp. 407–409.

compulsive belief to be sound, under appropriate circumstances. Bemused with syntax, the early logical positivists were chary of the notion of truth; and then, bemused with Tarski's account of truth, analytic philosophers neglected to inquire how we come to believe or disbelieve simple propositions. Quite simply put, the point is: coming to have suitable degrees of belief in response to experience is a matter of training – a *skill* which we begin acquiring in early childhood, and are never quite done polishing. The skill consists not only in coming to have appropriate degrees of belief in appropriate propositions under paradigmatically good conditions of observation, but also in coming to have appropriate degrees of belief between zero and one when conditions are less than ideal.

Thus, in learning to use English color words correctly, a child not only learns to acquire degree of belief 1 in the proposition that the cloth is blue, when in bright sunlight he observes a piece of cloth of uniform hue, the hue being squarely in the middle of the blue interval of the color spectrum: he also learns to acquire appropriate degrees of belief between 0 and 1 in response to observation under bad lighting conditions, and when the hue is near one or the other end of the blue region. Furthermore, his understanding of the English color words will not be complete until he understands, in effect, that blue is between green and violet in the color spectrum: his understanding of this point or his lack of it will be evinced in the sorts of mistakes he does and does not make, e.g. in mistaking green for violet he may be evincing confusion between the meanings of 'blue' and of 'violet', in the sense that his mistake is linguistic, not perceptual.

Clearly, the borderline between factual and linguistic error becomes cloudy, here: but cloudy in a perfectly realistic way, corresponding to the intimate connection between the ways in which we experience the world and the ways in which we speak. It is for this sort of reason that having the right language can be as important as (and can be in part identical with) having the right theory.

Then learning to use a language properly is in large part like learning such skills as riding bicycles and flying aeroplanes. One

must train oneself to have the right sorts of responses to various sorts of experiences, where the responses are degrees of belief in propositions. This may, but need not, show itself in willingness to utter or assent to corresponding sentences. Need not, because e.g. my cat is quite capable of showing that it thinks it is about to be fed, just as it is capable of showing what its preference ranking is, for hamburger, tuna fish, and oat meal, without saying or understanding a word. With people as with cats, evidence for belief and preference is behavioral; and speech is far from exhausting behavior [1].

Our degrees of beliefs in various propositions are determined jointly by our training and our experience, in complicated ways that I cannot hope to describe. And similarly for conditional subjective probabilities, which are certain ratios of degrees of belief: to some extent, these are what they are because of our training – because we speak the languages we speak. And to this extent, conditional subjective probabilities reflect *meanings*. And in this sense, there can be a theory of degree of confirmation which is based on analysis of meanings of sentences. Confirmation theory is therefore semantical and, if you like, logical [2].

---

[1] Jeffrey, *op. cit.*, pp. 57–59.

[2] Support of U.S. Air Force Office of Scientific Research is acknowledged, under Grant AF–AFOSR–529–65.

DISCUSSION

L. HURWICZ: *Richard Jeffrey on the three prisoners.*

I would like to make a comment which I think is along the lines of Professor Jeffrey's discussion of the three prisoners. I would like to make the situation a little more explicit than it was earlier, although I shall not contradict anything that has been said: I think this will help us to see to what extent, if any, there is anything surprising or paradoxical in the situation.

First of all let me say this: there were three possible decisions by the warden – AB, BC and AC; then, as against that, there was also the question of what the warden would say to a who asked the question who else was being freed, and clearly the warden could only answer 'b' or 'c'. What I'm going to put down here is simply the bivariate or two-way probability distribution, and it doesn't matter at all at this stage whether we interpret it as a frequency or as a subjective probability, because it's just a matter of applying the mechanics of the Bayes theorem.

One other remark I'd like to make is this: the case that was considered by Professor Jeffrey was one where the *a priori* probabilities of AB, BC and AC were each one-third. This actually does not at all affect the reasoning, and I will stick with it just because it is close to my limitations in arithmetic.

So the marginal frequencies or probabilities are all equal to one-third. If the decision had been AB, then of course the warden could only answer 'b', and similarly if the decision had been AC, he could only answer 'c'. So the joint frequency or probability of the following event is one-third: the people chosen for freeing are a and b, and when the warden is asked, 'Who is the person other than a who is about to be freed?', his answer is 'b'. The joint probability is also one-third that the choice was AC and that the warden answered 'c'.

We now come to the only case where the warden has a choice of what he will say, namely, the case where the decision was BC.

The question was raised, quite properly, of how he goes about making this choice.

Let me here say the following. In a sense what I'm doing here is a sally into enemy territory, because I personally am not particularly Bayesian in my approach to decision theory, so I would not myself assert that the only method is to describe the warden's decision, the warden's principle of choice, as a probabilistic one. However, if it is not probabilistic, then of course the prisoner, our a, would have to be using some other principle of choice on his part in order to decide what to do. Being an unrepentant conservative on this, I might choose, or A might choose, the minimax principle. However, in order to follow the spirit of the discussion here, I will assume that the whole thing is being done in a completely Bayesian or probabilistic way; in this case, to compute the remaining joint distribution we must make some probabilistic assumption about how the warden will behave when asked the question.

So let the principle be this, that he has a certain random device such that if the people to be freed are b and c, his answer to the question will be 'b' with probability β and 'c' with of course probability $1-\beta$. All I will assume for the moment about β is that it is between zero and one, and that's probably one of the few uncontroversial points so far.

It is clear that the sum of the two joint probabilities (BC and 'b', and BC and 'c') will be one-third; so the first will be $\frac{1}{3}\beta$, and the second $\frac{1}{3}(1-\beta)$. The marginal (or absolute) probabilities of 'b' and 'c' will be $\frac{1}{3}(1+\beta)$ and $\frac{1}{3}(2-\beta)$ respectively.

Inf. → / Dec. ↓	'b'	'c'	Marginal
AB	$\frac{1}{3}$	0	$\frac{1}{3}$
BC	$\beta/3$	$(1-\beta)/3$	$\frac{1}{3}$
AC	0	$\frac{1}{3}$	$\frac{1}{3}$
Marginal	$(1+\beta)/3$	$(2-\beta)/3$	1

Now what are the probabilities after the warden has given his answer? Suppose that the answer that the warden gave is 'b': the problem now is, what is the probability that a is to be freed, given that the warden said that b is to be freed? This probability, which I will denote by 'π_b', I obtain in the following way using what I hope is a self-explanatory notation:

$$\begin{aligned}\pi_b &= p(A/\text{`}b\text{'}) \\ &= \frac{p(A\cdot\text{`}b\text{'})}{p(\text{`}b\text{'})} \\ &= \frac{p(AB\cdot\text{`}b\text{'})+p(AC\cdot\text{`}b\text{'})}{p(AB\cdot\text{`}b\text{'})+p(AC\cdot\text{`}b\text{'})+p(BC\cdot\text{`}b\text{'})} \\ &= \frac{\frac{1}{3}+0}{\frac{1}{3}+0+\beta/3} \\ &= 1/(1+\beta).\end{aligned}$$

Similarly I get π_c (the probability that a is to be freed, given that the warden said that b is to be freed) as follows:

$$\begin{aligned}\pi_c &= p(A/\text{`}c\text{'}) \\ &= \frac{p(A\cdot\text{`}c\text{'})}{p(\text{`}c\text{'})} \\ &= \frac{p(AB\cdot\text{`}c\text{'})+p(AC\cdot\text{`}c\text{'})}{p(AB\cdot\text{`}c\text{'})+p(AC\cdot\text{`}c\text{'})+p(BC\cdot\text{`}c\text{'})} \\ &= \frac{0+\frac{1}{3}}{0+\frac{1}{3}+(1-\beta)/3} \\ &= 1/(2-\beta).\end{aligned}$$

Now the question which we now have to ask is this: are these conditional probabilities, π, different from the marginal (absolute) probability that a is to be freed, $p(a)$? And the answer is that they are except when β happens to be equal to one-half, in which case the probability remains at its marginal value of two-thirds. But except in this special case the probabilities π_b and π_c can vary from one-half to one [1].

[1] In the problem as reported by Mosteller, it might be reasonable to take $\beta=\frac{1}{2}$. In that case, let us note, $\pi_b=1/(1+\frac{1}{2})=\frac{2}{3}$ (not $\frac{1}{2}$ as suggested

As I indicated before, there is no quarrel between us, but I do want to explore just one step further, and that is this. You remember when we were told this anecdote there was a wave of laughter and I now want to see what it was that was so funny. It is that this prisoner became doubtful about asking for this extra information, because he thought his probability of being released would go down after getting it. So it seemed that having this extra information would make him less happy, even though he didn't have to pay for it. That really was the paradox, not the fact that the probabilities changed. Clearly, the change in probabilities is itself not at all surprising; for example, if the warden had told a the names of *two* people other than himself who would be freed, his optimism would have gone down very drastically [1].

What is surprising is that a thought he would be less happy with the *prospect* of having the extra piece of information than without this prospect. What I want to show now is that a was just wrong to think this; in other words, if this information was free, he should have been prepared to hear it.

Suppose for instance β is different from one-half: I think it is implicit in this little anecdote that the probability of a's being released either before or after getting the information, in some sense corresponds to his level of satisfaction. If his chances are good he is happy; if his chances are bad he is miserable. So these π's, though they happen to have been obtained as probabilities, may at the same time be interpreted as utilities or what Professor Jeffrey called desirabilities. Good. Now if a proceeds in the Bayesian way he has to do the following: he has to look at all these numbers, because before he asks for the information he does not know whether the answer will be 'b' or 'c'. Then he must ask himself the following: How happy will I be if he says 'b'? How happy will I be if he says 'c'? And then in the Bayesian (or de Finetti or Ramsey) spirit he multiplies the utilities, say $u(\text{`}b\text{'})$ and $u(\text{`}c\text{'})$ associated with hearing the warden say 'b' or 'c' by the respective probabilities, say $p(\text{`}b\text{'})$ and $p(\text{`}c\text{'})$, of hearing these answers. He thus obtains an expression for his expected [1] utility associated with getting the extra information, say

$$Eu^* = p(\text{`}b\text{'}) \cdot u(\text{`}b\text{'}) + p(\text{`}c\text{'}) \cdot u(\text{`}c\text{'}).$$

Now the required probabilities are the marginal probabilities at the bottom of the table, i.e.,

$$p(\text{`}b\text{'}) = \frac{1+\beta}{3}, \quad p(\text{`}c\text{'}) = \frac{2-\beta}{3}.$$

As for utilities, it is implicit in the argument that they are linear [2] functions of the probabilities that a will be released given the warden's answer. So

$$u(\text{`}b\text{'}) = \pi_b = \frac{1}{1+\beta}, \quad u(\text{`}c\text{'}) = \pi_c = \frac{1}{2-\beta}.$$

Hence the (expected) utility associated with getting the extra information from the warden is

$$Eu^* = \frac{1+\beta}{3} \cdot \frac{1}{1+\beta} + \frac{2-\beta}{3} \cdot \frac{1}{2-\beta} = \tfrac{2}{3}.$$

On the other hand, the expected utility Eu°, associated with *not* asking the warden for extra information is simply equal to the original probability $p(a)$ that a will be released,

$$Eu^\circ = p(a) = \tfrac{2}{3}.$$

---

[1] In the sense of the mathematical expectation of a random variable.

[2] See footnote 2 on the next page.

---

in the statement of the problem!) and also $\pi_c = 1/(2-\beta) = 1/(2-\frac{1}{2}) = \frac{2}{3}$. Hence (for $\beta = \frac{1}{2}$) a was wrong to expect the probabilities to change. But, on the other hand, the warden's reply would give him no additional information.

[1] Or suppose, that $\beta = 1$ (and a knows this). Then if a hears the warden tell him that c is one of the persons to be released, he will have good reason to feel happy. For when $\beta = 1$, the warden will tell a about having selected c only if the selected pair was AC. On the other hand, still with $\beta = 1$, if the warden says that b is one of the persons to be released, this means (with equal probabilities) that either AB or BC has been chosen, but *not* AC. Hence, with the latter piece of information, a will be justifiably less optimistic about his chances of release. (With β close to one, a similar situation prevails.)

Hence it so happens that (for a utility function linear in probabilities of release)

$$Eu^* = Eu^\circ,$$

i.e., the expected utility with extra information (Eu^*) is the same as without extra information (Eu°). Thus a should be willing (but not eager) to ask for extra information (if it is free of charge). 'On the average' [1], it won't do him any harm; nor will it help him [2].

P. SUPPES: *Rational changes of belief.*

I am generally very much in agreement with Professor Jeffrey's viewpoint on belief and knowledge as expressed in his paper. The focus of my brief remarks is to point out how central and difficult are the problems concerning *changes* of belief. Jeffrey remarks that the familiar method of changing probable beliefs by explicitly conditionalizing the relevant probability measure is not adequate in many situations – in fact, in all those situations that involve a change in the probability assigned to evidence, but a change that does not make the probability of possible evidence 0 or 1.

My point is that once we acknowledge this fact about the probable character of evidence we open Pandora's box of puzzles for any theory of rational behavior. I would like to mention three problems. These problems are not dealt with explicitly by Jeffrey, but the focus I place on them is certainly consistent with his own expressed views.

1. *Attention and selection.* I begin with the problem of characterizing how a man attempting to behave rationally is to go about selecting and attending to what seems to be the right kind of evidence. Formulations of the problem of evidence within a logically well-defined language or sample space have already passed over

[1] 'On the average' expresses the fact that the decision is made on the basis of mathematical expectation. It need not imply a frequency interpretation of probabilities.

[2] When utilities are *non-linear* with respect to probabilities of release, the prospect of additional information may be helpful or harmful.

the problem of evaluating their appropriateness. Any man has highly limited capacities for attention and observation. Given a characterization of his powers and limitations what is a rational way to behave? Consider the familiar coin-flipping example. When is it appropriate to stop concentrating on the outcomes of the flips and to start observing closely the behavior of the gambler doing the flipping? To take another sort of example, how do we characterize the behavior of detectives who act on subthreshold cues and conjectures that can scarcely even be verbalized, let alone supported by explicit evidence? Put more generally, what is the rational way to go about discovering facts as opposed to verifying them? It is easy to claim that for a wide variety of situations rational discovery is considerably more important than rational verification. In these cases of discovery we need to understand much better the information-processing capacities of human beings in order to be able to characterize in a serious way *rational* information-processing. Above all, it is certainly not clear to me what proportion of rational information-processing should be verbalized or is even verbalizable.

2. *Finite memory.* The second problem concerns the rational use of our inevitably restricted memory capacities. A full-blown theory of rationality should furnish guidelines for the kinds of things that should be remembered and the kind that should not. Again a solution, but certainly not a solution whose optimality can be seriously defended, is at least partly given by the choice of a language or a sample space for dealing with a given set of phenomena. But the amount of information impinging on any man in a day or a week or a month is phenomenal and what is accessible if he chooses to make it so is even more so. What tiny fraction of this vast potential should be absorbed and stored for ready access and use? Within the highly limited context of mathematical statistics, certain answers have been proposed. For example, information about the outcome of an experiment can be stored efficiently and with little loss of information in the form of the likelihood function or some other sufficient statistic, but this approach is not of much use in most situations, although elements of the approach

can perhaps be generalized to less restricted circumstances. Perhaps even more importantly, it is not clear what logical structure is the most rational to impose on memory. The attempts at constructing associative computer memories, as they are often called, show how little we are able as yet to characterize explicitly a memory with the power and flexibility of a normal man's, not to speak of the memory of a normal man who is using his powers with utmost efficiency. Perhaps one of the most important weaknesses of confirmation theory and the Ramsey-sort of theory developed by Jeffrey and others is that little is said about imposing a logical structure on evidence. Part of the reason for this is that the treatment of evidence is fundamentally static rather than dynamic and temporal. In real life, evidence accumulates over time and we tend to pay more attention to later than earlier data, but the appropriate logical mechanisms for storing, organizing and compressing temporally ordered data are as yet far from being understood.

3. *Concept formation.* The most fundamental and the most far-reaching cognitive changes in belief undoubtedly take place when a new concept is introduced. The history of science and technology is replete with examples ranging from the wheel to the computer, and from arithmetic to quantum mechanics. Perhaps the deepest problem of rational behavior, at least from a cognitive or epistemological standpoint, is to characterize when a man should turn from using the concepts he has available to solve a given problem to the search not just for new evidence but for new concepts with which to analyze the evidence. Perhaps the best current example of the difficulty of characterizing the kinds of concepts we apply to the solution of problems is the floundering and vain searching as yet typical of the literature on artificial intelligence. We cannot program a computer to think conceptually because we do not understand how men think conceptually, and the problem seems too difficult to conceive of highly nonhuman approaches. For those of us interested in rational behavior the lesson to be learned from the tantalizing yet unsatisfactory literature on artificial intelligence is that we are a long way from being able to say what a rational set of concepts for dealing with a given body of experience should

be like, for we do not have a clear idea of what conceptual apparatus we actually use in any real sense.

To the problems about rationality I have raised in these remarks there is the pat answer that these are not problems of the theory of rational behavior but only of the theory of actual behavior. This I deny. A theory of rationality that does not take account of the specific human powers and limitations of attention, memory and conceptualization may have interesting things to say but not about human rationality.

R. C. JEFFREY: *Reply.*

Suppes' and Hurwicz' comments are interesting and instructive, and I find I have little to add to them. But perhaps a brief postscript is in order, in response to Suppes' closing remark:

> A theory of rationality that does not take account of the specific human powers and limitations of attention may have interesting things to say, but not about human rationality.

It may be that there is no real issue between us here, but the emphasis makes me uncomfortable. In my view, the logic of partial belief is a branch of decision theory, and I take decision theory to have the same sort of relevance to human rationality that (say) quantification theory has: the relevance is there, even though neither theory is directly about human rationality, and neither theory takes any account of the specific powers and limitations of human beings.

For definiteness, consider the following preference ranking of four sentences s, s', t, t', where s and s' are logically inconsistent, as are t and t'.

$$s$$
$$s'$$
$$t$$
$$t'$$

This ranking is *incoherent*: it violates at least one of the following two requirements. (a) Logically equivalent sentences are ranked

together. (b) The disjunction of two logically incompatible sentences is ranked somewhere in the interval between them, endpoints included. Requirements (a) and (b) are part of (or anyway, implied by) a definition of 'incoherent'. To see that the given ranking is incoherent, notice that (a) implies that the disjunction of the sentences s, s' is ranked with the disjunction of the sentences t, t', while (b) implies that in the given ranking, the first disjunction is higher than the second. In my view, the point of classifying this ranking as incoherent is much like the point of classifying the pair s, s' as logically inconsistent: the two classifications have the same sort of relevance to human rationality. In the two cases, a rational man who made the classification would therefore decline to own the incoherent preference ranking or to believe both of the inconsistent sentences. (For simplicity I speak of belief here as an all-or-none affair.)

True enough: since there is no effective decision procedure for quantificational consistency there is no routine procedure a man can use – be he ever so rational – to correctly classify arbitrary rankings of sentences as incoherent or arbitrary sets of sentences as inconsistent. The relevance of incoherence and inconsistency to human rationality is rather that a rational man, once he comes to see that his preferences are incoherent or that his beliefs are inconsistent, will proceed to revise them. In carrying out the revision he may use decision theory or quantification theory as an aid; but neither theory fully determines how the revision shall go.

In fine, I take Bayesian decision theory to comprise a sort of *logic* of decision: the notion of coherence has much the same sort of relationship to human ideals of rationality that the notion of consistency has. But this is not to deny Suppes' point. The Bayesian theory is rather like a book of rules for chess which tells the reader what constitutes winning: there remains the question of ways and means.

INDUCTION BY ENUMERATION AND INDUCTION BY ELIMINATION

JAAKKO HINTIKKA

Stanford University and University of Helsinki

The most striking general feature of the theory of induction at the present moment is the division of the field into several different schools of thought [1]. The best known contrast between different points of view is undoubtedly the difference between the views of Professor Carnap and of Sir Karl Popper [2]. There are other discrepancies between different approaches, however, which are not unrelated to the Carnap–Popper debate but which also have other ramifications and other sources. A well-known case in point is the contrast between those theories of induction in which some sort of rule of acceptance is used or argued to be unavoidable and those in which such rules play no part and in which one is instead supposed to be content with assigning different degrees of probability (on evidence) to the statements one considers [3]. There is

[1] For general surveys, see Henry E. Kyburg, 'Recent work in inductive logic', *American Philosophical Quarterly* 1, 1964, pp. 1–39, and S. F. Barker, *Induction and Hypothesis*, Cornell University Press, Ithaca, New York, 1957.

[2] The basic works are: Rudolf Carnap, *Logical Foundations of Probability*, University of Chicago Press, Chicago, 1950, 2nd ed., 1962, and Karl R. Popper, *The Logic of Scientific Discovery*, Hutchinson and Co., London, 1959. See also Karl R. Popper, *Conjectures and Refutations*, Routledge and Kegan Paul, London, 1963; Rudolf Carnap, *The Continuum of Inductive Methods*, University of Chicago Press, Chicago, 1952, and the relevant parts of P. A. Schilpp, ed., *The Philosophy of Rudolf Carnap*, The Library of Living Philosophers, Open Court Publishing Company, La Salle, Illinois, 1963.

[3] Cf. e.g. Isaac Levi, 'Corroboration and rules of acceptance', *The British Journal for the Philosophy of Science* 13, 1963, pp. 307–313, and Henry E. Kyburg, 'Probability, rationality, and a rule of detachment', *Logic, Methodology and Philosophy of Science, Proceedings of the 1964 International Congress for Logic, Methodology and Philosophy of Science*, Y. Bar-Hillel ed., North-Holland Publishing Company, Amsterdam, 1965, pp. 301–310.