Richmond Campbell

Review Article

Rational Decision and Causality. *By Ellery Eells. Cambridge and New York: Cambridge University Press, 1982, Pp. x, 234. $29.50 U.S.*

The book begins with a compact but comprehensive introduction to the Bayesian theory of rational decision and concludes with an elaborate explanation of its implications for Newcomb decision problems. In the middle is a useful survey of the different approaches of Ramsey, Savage, and Jeffrey, and the causal decision theorists, Gibbard, Harper, Skyrms, and Lewis. The discussion throughout is informative, technically expert, and fair. There are also some worthwhile digressions. For example, philosophers interested in inductive reasoning would do well to study Eells' illuminating application of Bayes' Theorem to the Raven paradox, Goodman's riddle about "grue", and various other puzzles regarding the logic of confirmation. I have difficulty, however, with Eells' answer to Newcomb's problem. My review will focus on that issue after some general observations about Bayesian decision theory.[1]

The central idea of that theory may appear self-evident: A decision is rational, relative to the agent's beliefs and desires, if and only if its probable outcome is at least as desirable as the probable outcome of any alternative decision that is possible for the agent in the circumstances. What could be more obvious? One might be tempted to say that this conception of rational choice is presupposed in *every* explanation of deliberate action. (Take any psychoanalytic explanation of an action that is unintelligible relative to conventional desires and beliefs. The action is intelligible relative to the unconscious desires and beliefs postulated in the explanation, one could argue, because relative to them the action is rational within the Bayesian model.) But one might worry that this conception of rationality is ultimately without empirical content. One's doubts will be strengthened when one learns that the theory places very little restriction on what is to count as desirable or probable from the point of view of the decision-maker. A "rational" person's desires and beliefs about what is probable can be as outlandish as you please provided only that they meet certain minimal constraints of internal consistency. Furthermore, the existence of those desires and beliefs is to be inferred from the decision-maker's actions. Small wonder that the Bayesian norm appears self-evident! If we are prepared to believe that someone's actions are rational—relative to what we can infer to be desirable and probable for that person—what evidence could possibly convince us that this norm has been violated?

Bayesians have not been insensitive to the problem of circularity. In a standard mathematical formulation the Bayesian norm says that a decision is rational if it maximizes conditional expected utility. The conditional expected utility of an action is the average degree of desirability of its possible outcomes weighted by the subjective probabilities of the outcomes conditional on the performance of the action. Much of the mathematical development of Bayesianism in the last sixty years has been directed toward defining appropriate measures for degrees of subjective desirability and probability. (The latter are roughly degrees of belief.) Various methods of measurement have been proposed. For example, following Ramsey, one can begin with minimal assumptions about a person's desirability function and then on the basis of the person's choices in simple lotteries construct a subjective probability function; or, following von Neumann and Morgenstern, one can start with some assumptions about subjective probability and work towards a desirability function. In both cases, the desire and belief functions are interdependent and based on the agent's choice behaviour, but the behavioural implications that can be derived reach far beyond the behavioural data on which desire and belief functions are based. Discrepancies between a person's actual choices and the choices required by the Bayesian norm can emerge when these functions are only partly specified. Thus, despite its apparent triviality when expressed in terms of common sense, modern Bayesianism, in its various refined formulations, is an empirically testable theory.

Ironically, the real problem is just the opposite. Various kinds of evidence, the Allais and Ellsberg paradoxes, the Kahneman-Tversky experiments, and Newcomb's problem, suggest that persons whom we would otherwise judge to be clear-headed, balanced, and well-informed—persons who seem otherwise "rational"—choose contrary to Bayesian rationality when faced with certain decision problems. Not that this evidence by itself is generally regarded as a refutation of Bayesianism. The theory is sufficiently complex to allow possibilities for adjustment (for example, the introduction of a "regret" factor to explain away the Allais and Ellsberg paradoxes, pp. 39-40). That the discrepancies are known as "paradoxes" and "problems" suggests the continuing strength of the Bayesian paradigm. Still, there should be no doubt that there exists *prima facie* evidence against the theory.[2]

A possible response is to argue that the evidence shows only that people are not rational. One might argue, in other words, that the theory can be demonstrated to be fundamentally mistaken about actual human motivation without being refuted as a *normative* ideal. To his credit Eells does not make this familiar move. By implication he allows that the theory might be fundamentally mistaken in both respects. His position is (pp. 8-9) that subjective desirability and probability functions are best construed as theoretical entites postulated to explain behaviour within a certain theoretical context (defined by the Kolmogorov axioms of probability, the Bolker representation theorems, and so on). On this view it would be possible to conclude, if evidence is sufficiently recalcitrant, that these functions don't exist, anymore than phlogiston does. If Bayesianism is *that* mistaken about actual motivation, then how is anyone to embody its ideal for rational motivation? That would not be possible even in principle. To defend Bayesianism as an ideal is to imply that it provides a description of actual motivation that is at least approximately true. There is also a more immediate objection to retreating from the descriptive to the prescriptive. In Newcomb decision problems people sometimes fail to choose according to the Bayesian ideal because they believe that it would be *irrational* to do so.

A lucid description of Newcomb's Problem is provided by Howard Sobel in "Predicted Choices" (this volume, pp. 600-607). Here is a variation invented by Sobel.[3] One thousand dollars is put before you, to take or leave, with no strings attached, *except* that you know that someone has already deposited one million dollars to your bank account if and only if that person has predicted that you will not take the thousand. Assume that the relative subjective desirability of each outcome is represented in the matrix below.

		Predictor's Deposit	
		$M	$0
Your action	Take $1000	1001	1
	Leave $1000	1000	0

The numbers in the cells indicate proportional desirability. Leaving the thousand and getting the million is for you a thousand times better than having the thousand alone and only slightly inferior to having a million plus a thousand. But either the deposit has been made or it has not been made. Nothing you do *now* can change what the predictor has already done. Consequently, when you make your choice, you will be choosing, in effect, between outcomes in a single column of the above matrix, either between 1001 and 1000 (in case the $M is already in your account) or between 1 and 0 (in case nothing has been deposited). To many people it appears obvious that the rational choice is to take the thousand dollars, no matter how sure you are that the predictor has predicted correctly. They reason: You will be one thousand dollars ahead, whether the million is there or not. Sobel would agree: taking the "extra" money is the rational choice.

But Bayesianism *seems* to lead to just the opposite conclusion. For the purpose of illustration suppose the evidence of the predictor's past success is such that your conditional subjective probability function has the following values.

		Predictor's Deposit	
		$M	$0
Your action	Take $1000	.01	.99
	Leave $1000	.99	.01

The fractions in the cells may be regarded as showing your degree of confidence that the deposit is as indicated on the assumption you perform the action indicated. The Bayesian rational choice is the action with the highest expected value, calculated from a weighted average of the proportional desirabilities given in the first matrix.

$$\text{Exp (take)} = (.01 \times 1001) + (.99 \times 1) = 11$$

$$\text{Exp (leave)} = (.99 \times 1000) + (.01 \times 0) = 990$$

Thus Bayesianism apparently implies that the rational decision is to *leave* the thousand dollars.

Causal decision theorists argue that the trouble lies in using conditional probabilities to calculate a weighted average of the value of the possible outcomes of an action. In many cases this method of averaging will give the right answer because conditional probabilities often reflect the degree of probable causal relevance of the action for the outcome. But, in other cases, as in decision problems with the above structure, the probabilistic dependence of the relevant states of the world on the agent's action is misleading. These states, though probabilistically dependent on the actions, are *causally independent* of them. Causal decision theorists maintain that a rational person should be concerned only about the actual effects that a decision is believed to cause. These theorists therefore propose to modify the standard theory. Like traditional Bayesians they identify rationality with maximizing expected utility, but they think of expected utility differently. Instead of using conditional probabilities to calculate the expected value of an action, they formulate a weighting system designed to reflect the degree of probable *causal* dependence of the relevant states on the action. There are disagreements among them about how to formulate this measure. All agree, however, that it will diverge from conditional subjective probability in many important cases.

There is some danger of terminological confusion here. Causal decision theorists could be regarded as Bayesians in that they identify rational action with "maximizing expected utility" and define the latter in terms of certain *subjective* desirability and probability functions. They are, however, importantly different from traditional Bayesians. For convenience I shall reserve "Bayesianism" for the traditional variety.

Eells' defense of Bayesianism has two parts. He argues first that its causal counterpart is burdened with causal concepts, and these are, as every philosopher knows, difficult to make clear. It is not as if causal decision theory trades probabilistic concepts for causal ones. Causal decision theory has both; all versions of the causal theory require a subjective probability distribution over various competing causal hypotheses. Bayesianism is simpler, and, other things being equal, the simpler theory is to be preferred. Second, Eells argues that other things *are* equal because the alleged counterexamples constructed from Newcomb problems are illusory. He contends that traditional Bayesianism, properly applied, recommends the *same* action as causal decision theory.

How is this possible? The answer is a bit involved. If we are to take Newcomb's Problem seriously, we have to suppose that there is some explanation of the predictor's amazing ability to predict choices.[4] Presumably there is something about the agent, perhaps a genetic condition recognizable to the predictor (it doesn't matter what), that is causally responsible for both the predictor's prediction and the agent's choice. There must be, that is to say, *a common cause* that explains the probabilistic dependence in the absence of any direct causal influence of the choice on the prediction. This supposition will bother those who think that a rational action can not have this kind of causal ancestry but we cannot stop to pursue that issue.[5]

Label the choice of taking the thousand dollars S and not taking it $\overline{S}$. (Think of S standing for "seizing" the extra money). Let G be the genetic condition (or whatever) that is responsible for S and $\overline{G}$ be its absence. On the assumption that *S* is to be explained almost entirely by *G*, we can represent the Newcomb problem as follows:

Desirabilities

	$\overline{G}$	G
S	1001	1
$\overline{S}$	1000	0

Probabilities

	$\overline{G}$	G
S	.01	.99
$\overline{S}$	.99	.01

Various stories will fit this structure. Imagine *G* to be a genetic condition that is the common cause of smoking, *S*, and lung cancer (R.A. Fisher's hypothesis[6]). Or *S* might be your living a life of sin and *G* God's having preordained that you burn in hell.[7] Another example, much discussed in the literature[8], is the notorious Prisoner's Dilemma. Imagine you are in a one play Prisoner's Dilemma with someone who you believe is psychologically your twin and has precisely the same information. Let *S* be your "selfish" choice of non-cooperation and *G* be whatever feature of the choice situation would generate non-cooperation in a person psychologically like you. You are nearly certain that if you choose *S*, then the *G* factor is present. By comparison with mutual cooperation the outcome in that case would be terrible. (Look at the desirability matrix!) The critical feature that makes these Newcomb problems is that the conditional probabilities are explained by *G* causing *S* rather than the reverse. This is also the key to Eells' defence.

To block the objection against Bayesianism, Eells needs to show that *S* is rational on this theory, despite the initial conditional probabilities. He tries to do this by using the common cause structure to develop a variation on the so-called "tickle defence". A tickle defense would run as follows. The genetic condition (or whatever) that causes *S* cannot do so except through the agent's psychological states. Imagine that the last psychological state mediating the causal connection between *G* and *S* can be introspected. It is like having a tickle before you seize the thousand dollars (or smoke or sin or act selfishly). Label it *T* and let *P(X/ Y)* be the conditional probability of *X* on *Y*. Since *T* carries all the causal information contained in *G* regarding its effect on *S*, we can expect *T* to "screen off" the probabilistic relevance of *G* for *S*.[9]

$$P\,(S/\,G\&T) = P\,(S/\,\overline{G}\&T)$$

That is to say, *S* is just as probable, given *T*, whether or not *G* obtains. By the symmetry of probabilistic independence,

$$P\,(G/\,S\&T) = (P\,(G/\,\overline{S}\&T\,)$$

But since T is introspectible, the agent will be certain that T obtains. Thus, the subjective unconditional probability of T should be unity.

$$P(T) = 1$$

From the last two equations, it follows that for an agent who is immediately aware of the tickle and of its causal significance, the state G is probabilistically *independent* of action S:

$$P(G/S) = P(G/\overline{S})$$

It also follows, of course, that $\overline{G}$ is probabilistically independent of S.

$$P(\overline{G}/S) = P(\overline{G}/\overline{S})$$

From here it is a very short step to the conclusion that S maximizes conditional expected utility.

$$\text{Exp}(S) = (1001 \text{ x } P(\overline{G}/S)) + (1 \text{ x } P(G/S))$$

$$\text{Exp}(\overline{S}) = (1000 \text{ x } P(\overline{G}/\overline{S})) + (0 \text{ x } P(G/\overline{S}))$$

$$\text{Exp}(S) > \text{Exp}(\overline{S})$$

The objections that have been made to the tickle defence are formidable.[10] Suppose that I am advising my great aunt about whether she should smoke and I know that G is the common cause of S and lung cancer. If I know that her values are as shown in the desirability matrix, then I should advise her to smoke — and I am justified without having to know anything about her tickles. I know she is better off doing S whether or not T obtains. Put another way, I can sensibly advise her to do S, even though *my* probability function would assign her T an unconditional probability much less than one. This line of objection can be taken a step further. Why suppose my great aunt has any tickles that she can identify as the cause of S? Suppose *her* probability function is such that $P(T) = .01$. Even so, she should still be justified in smoking on the information she already has.

Suppose, however, that there is a tickle that every rational agent must have. A rational agent, one might argue, must act on reasons. These reasons would be the last causal link between G and S. Call the psychological state of having these reasons R. By substituting R for T in the previous tickle defence, we have again a Bayesian argument that S is the rational choice. This time it appears that the previous objections can be met. I can be sure that my great aunt has R and is aware of R if she is a rational agent. Thus, I can advise her how to choose — or she can decide for herself — by the Bayesian argument already given.

Since R must always exist and be accessible to a rational agent, the previous objections are blocked.

This is indeed Eells' defence stripped down to its bare essentials. (Or at least I presume to say it is. The last paragraph paraphrases some fifty pages of careful argument in Eells.) My difficulty[11] is that the defence seems to rest on an equivocation. There are two possible references for R. It could be the last causal link between G and S *or* the last one between $\overline{G}$ and $\overline{S}$. Call the two possibilities $R(G)$ and $R(\overline{G})$, respectively. Remember that my great aunt, when she is about to carry out the Bayesian reasoning, doesn't yet know whether she will choose S or $\overline{S}$, and she doesn't know, of course, whether she has G or $\overline{G}$. If she could somehow know, *before* she completes her deliberation, whether she has $R(G)$ or $R(\overline{G})$, then she could certainly invoke the tickle defense. However, what guarantee is there that she will be able to know this? No part of being a rational person guarantees this particular knowledge, which involves knowing the causal ancestry of her reasons. At most she can be said to know that R will be either $R(G)$ or $R(\overline{G})$. But that is not enough to allow her to go through the steps of the tickle defense. In sum, there is no guarantee in this decision problem that a rational person will reach a rational conclusion by Bayesian reasoning.

May my great aunt not reason by constructive dilemma? Either R is $R(G)$ or it is $R(\overline{G})$. Either way, by the steps of the tickle defence, S is rational. Ergo, S is rational. In the jargon of decision theory, this is "dominance reasoning". It is precisely parallel to the argument of the causal decision theorist who argues: Either her genetic condition is G or it is $\overline{G}$. Either way the expected value of S is greater than that of $\overline{S}$. Ergo, S is rational. This style of reasoning is fine for a causal decision theorist, but it is legitimate for a Bayesian only if G, or in present instance, $R(G)$, is *probabilistically* independent of S. Manifestly $R(G)$ is not.

Are there any options left for a tickle defence of Bayesianism? In a recent article[12], Eells constructs a Bayesian account of the dynamics of deliberations (involving "metatickles"!) and argues that the last type of objection is met in this new account. These matters cannot be pursued here. Let me conclude, then, with this sentiment. There should be no doubt, on the evidence of the present work, that Bayesianism has in Eells an immensely knowledgeable and creative advocate. If he cannot reconcile the theory with the evidence, Bayesians should worry.

NOTES

[1] "Bayesianism" can mean either a theory of *learning* (based on Thomas Bayes' theorem about the probability of a hypothesis conditional on the evidence) or a theory of *decision*.

As Eells puts the difference: "just as a Bayesian *decision* theory tells you what course of action it is rational to pursue *relative* to your beliefs and desires, irrespective of how factually or morally justified they may be, so Bayesian *learning* theory tells you what new degree of belief assignment it is rational to adopt when new evidence comes in *relative* to what your prior degrees of belief are" (p. 12). In this review I shall mean primarily the Bayesian theory of rational decision, though the two theories are closely connected historically and conceptually.

[2] For a recent discussion of the negative implications of the Allais paradox, see Lanning Sowden, "The Inadequacy of Bayesian Decision Theory", *Philosophical Studies*, 45 (1984), 293-313.

[3] The first publication of William Newcomb's puzzle was in Robert Nozick, "Newcomb's Problem and Two Principles of Choice" in *Essays in Honor of Carl G. Hempel*, edited by Nicholas Rescher (Dordrecht: D. Reidel, 1969).

[4] Not all experts agree. James Cargile does not in his otherwise laudatory review of Eells' book in *The Journal of Philosophy*, 81 (1984), 163-168.

[5] This kind of issue is addressed in J.L. Mackie, "Newcomb's Paradox and the Direction of Causation", *Canadian Journal of Philosophy*, 7 (1977), 213-225.

[6] For a discussion of the relevance of this hypothesis for Newcomb's Problem, see Issac Levi, "Common Causes, Smoking, and Lung Cancer" in *Paradoxes of Rationality and Cooperation, Prisoner's Dilemma and Newcomb's Problem*, edited by Richmond Campbell and Lanning Sowden (Vancouver: University of British Columbia Press, 1985).

[7] The theological relevance of Newcomb's Problem is discussed in Steven Brams, *Superior Beings; If They Exist How Would We Know?* (New York: Springer-Verlag, 1983).

[8] The connection between the Prisoner's Dilemma and Newcomb's Problem was noted first in Nozick, *op. cit.*; but see also: David Lewis, "Prisoner's Dilemma Is a Newcomb Problem", *Philosophy and Public Affairs*, 8 (1979), 235-240; and Howard Sobel, "Not Every Prisoner's Dilemma Is a Newcomb Problem", in Campbell and Sowden, *op. cit.*, where Lewis' article is reprinted.

[9] A rigorous specification of the conditions under which an element in a causal chain will screen off earlier elements is proposed in Ellery Eells and Elliott Sober, "Probabilistic Causality and the Question of Transitivity", *Philosophy of Science*, 50 (1983), 35-37.

[10] In this paragraph I am drawing on objections from Frank Jackson and Robert Pargetter, "Where the Tickle Defence Goes Wrong", *Australasian Journal of Philosophy*, 61 (1983), 295-299, reprinted in Campbell and Sowden, *op. cit.*, and Brian Skyrms, *Causal Necessity* (New Haven and London: Yale University Press, 1980), pp. 130-131.

[11] A similar objection is made in Paul Horwich, "Decision Theory in the Light of Newcomb's Problem", manuscript, MIT, and independently in Ann Levey, "Newcomb's Problem and Rational Choice", unpublished M.A. thesis, Dalhousie University, 1984.

[12] Ellery Eells, "Metatickles and the Dynamics of Deliberation' ', *Theory and Decision*, 17 (1984), 71-95.