RATIONALITY IN DECISION THEORY AND IN ETHICS

HILARY PUTNAM
Universidad de Harvard

We are haunted by a certain culturally accepted distinction between "science" and "ethics", but we are also haunted by another culturally accepted distinction, the distinction between "absolute" and "instrumental" values —in effect, the distinction between valuing and *engineering*. Kant himself was in the grip or this dichotomy when he insisted that all "imperatives" must be either "hypothetical" or "categorical". The assumption has always been that "hypothetical imperatives", statements about what one ought to do *if* one wants to attain a particular end, are unproblematic in exactly the way that scientific statements are thought to be unproblematic. My purpose in this lecture will be to show that this is wrong; that if we are in a position that seems troubling in "ethics", we are inexactly the same position in "engineering", that the hypothetical imperative is in the same situation as the categorical, that rationality is as difficult a thing to "explain" in both cases.

I shall begin by discussing the formal assumptions behind what is called "decision theory", because I think those assumptions themselves are more dubious than is usually recognized. Then, in the second part of the paper, I shall examine a case proposed by Peirce.

The axioms of rational preference theory

When I first encountered modern decision theory and rational preference theory in the classic work or von Neumann, which I read many years ago, I was awed —who

wasn't? Like many other people, I felt that von Neumann had succeeded in recovering everything that was sound in classical utility theory without assuming the dubious psychology upon which classical utility theory based itself. But very soon I began to have doubts —doubts I want to share with you today.

I shall not spend much time on the ingenious use of the notion of an *ideally rational gambler* which underlies von Neumann's way of getting a utility scale. Von Neumann imagines a gambler who can answer such questions as "Do you prefer a gamble which gives you a chance r of getting X and a chance l-r of getting Y, or a gamble which gives you a chance s of getting X' and a chance l-s of getting Y', where X, Y, X', Y' are themselves "commodity bundles" —combinations of things as different as a concert, a place to live,a friend, etc., and where ones choices are required to be *rational*, as defined by the axioms for "rational preference". And he proves a beautiful theorem which shows that, if your preferences are perfectly rational and defined on such "gambles", then it is possible to assign utilities to the individual commodities in a way which rationalizes all the bets you are willing to make. What I want to discuss is the notion of "rational preference" itself. (In the second part of this lecture I shall talk about the other great notion of decision theory —the notion of "subjective probability".)

In order to derive his result, von Neumann needs to assume, of course, certain axioms for "rational preference". These axioms imply that all choices —including what are intuitively choices of "incomparable" alternatives— can be rank ordered: any two "commodities" are either (a) unequal (one is preferred to the other), or (b) equal (I am "indifferent" as to which I shall get).

Of course, many students feel uncomfortable when they encounter these axioms. The student often wants to say that there is a *tertium quid,* that he may, sometimes in his life, have to choose between "incomparable" alternatives, and that choosing between "incomparable" alternatives is not the same thing as choosing between alternatives that are

"indifferent" from his point of view. But the economist, or whoever, challenges the student to give the distinction between *indifference* and *incomparability* any content. And here the student normally finds himself stuck.

I want to defend the intuitive sense that a real distinction has been wiped out. First, however, let us identify the relevant axiom.

I don't think anyone objects to the assumption that (for an ideally rational agent) *preference is transitive.* Thus, using "xPy" to symbolize "x is preferred to y", we have

$$(xPy \ \& \ yPz) \text{—} xPz$$

What the "rational preference" theorist is assuming is something much stronger, however. He is assuming that *both preference and its complement are transitive.* That is, he is assuming both the validity of the previous axiom and that.

$$(\sim xPy \ \& \ \sim yPz) \text{—} \sim xPz$$

(In words: "If x is *not* preferred to y and y is *not* preferred to z, then x is *not* preferred to z.)

These two axioms together justify the claim that "indifference is an equivalence relation" ("xIy" —the chooser is indifferent as between x and y— is defined by

$$xIy = df \sim xPy \ \& \ \sim yPx).$$

The properties of an equivalence relation —reflexivity, symmetry, and transitivity— follow from the above axioms together with the assumption of the irreflexivity of preference. The "work", however, as measured by the strain upon intuition is done by the second axiom: the transitivity of the complement of the preference relation. Why is it so difficult to reject this axiom?

The difficulty that one appears to get into if one rejects this axiom can be easily sketched. Suppose I prefer x to y

and I claim that a third "commodity" z is indifferent[1] to both x and y, in the technical sense that zIx & zIy. Then a decision theorist can, it seems, convict me of being "irrational" in my expressed preferences by the following argument (analogous to a "Dutch book" argument[2] in probability theory): "Suppose," he says, "I were to offer you a choice between x and y. Since you prefer x to y, you would choose x.But, suppose, instead, you are confronted with the alternatives of x and z. Since you are indifferent, you cannot complain if I, instead of offering you a choice, just give you z rather than x. If you complain, that would show that after all, you did prefer x to z, contrary to your expressed statement that $\sim xPz$ & $\sim zPx$. Isn't that right?" Having gotten you to agree, he goes on, "But now, having gotten you to agree that it's all right if I give you z, I can say 'Since you don't care whether you get z or y, and it's turned out to be inconvenient for me to give you z after all, I will give you y instead'. If you complain at *this* stage, that will show you did prefer y to z, contrary to your expressed statement that $\sim yPz$ & $\sim zPy$. But if you don't, then in two steps I will have 'moved' you —with your consent at each step— from receiving x to receiving y —that is, from a preferred to a less preferred alternative".

What I have to do, to make good on my program of defending the intuitive objection to the assumption that the relation I is transitive, is defuse this argument. But this is not easy to do —not because the argument is invulnerable, but because it rests on nothing less than a whole way of thinking, and it is necessary to expose that way of thinking, and not just to think of a "counterexample" to an axiom.

A counterexample is, nonetheless, needed. And here is a simple one: suppose I am torn, as Pascal imagined me to be

[1] I say "x is indifferent to y" as short for "the chooser is indifferent as between x and y."

[2] A "Dutch book" is a system of bets that cannot result in a favorable outcome for the bettor, no matter how the gambles turn out. The axioms of subjective probability theory are frequently justified by proving that violation of them always leads to the possibility of being required to accept a "Dutch book."

in his famous "wager", between an ascetic-religious way of life and a hedonistic-sensual way of life. I may be quite sure that if I choose the hedonistic-sensual may of life, I would prefer to have a beautiful and responsive lover to a plain and unresponsive one. Call these choices x and y, and let z be the ascetic-religious life. If I regard the two ways of life as "incomparable", then I might insist that, prior to my making my existential choice, $\sim xPz \,\&\, \sim zPx$, and also $\sim yPz \,\&\, \sim zPy$ (*i.e.*, xIz and yIz). This is certainly the kind of case the student has in mind when he calls for a distinction between "incomparability" of alternatives and *mere* indifference. Why is the student not convicted of irrationality by the argument I just described, the argument that is so analogous to a Dutch book argument?

The problem with the "Dutch book" type argument just described is very simple: it ignores the *one* value that cannot itself be represented as just one more "commodity" to be combined with the various "bundles" among which or between which I am to choose: the value Kant called *autonomy*, the value of *making the choice myself as opposed to having it made for me.* It is part of calling the choices between x and z and between y and z *choices between ways of life,* that I view them as choices to be made by *me* in the process of *deciding who I am to be.* Regarding all choices as choices between *external* goods, goods that someone else may alot to me provided he respects my subject value assignments, is precisely the heart of the bureaucratic-managerial outlook that underlies the whole subject of decision theory. (Of course, reading xIy as "indifference" helps to conceal what is at stake). If someone "decides" to give me x rather than z on the grounds that I do not (yet) have a formed preference between these two ways of life, he deprives me of preciselly what is *most* important to me: namely, that the decision, whichever it is, shall be my own.

Could one defend rational preference theory by assuming that an ideally rational agent will already have *made* all his existential choices? To have made all possible existential choices is precisely to have stopped growing, to have become

utterly *rigid* as a human being. I cannot believe that anyone would really want to pack *this* attribute of some human personalities into "rationality"!

Peirce's example[3]

The example I wish to discuss is one that Peirce used to draw a certain connection between scientific problems and ethical problems —though not the one I would draw. In my opinion, Peirce's great contribution lies in his perception of the depth of individual problems, even if he did not succeed in building a unified system out of all those wonderful perceptions. One of these great flashes of genius occurs when Peirce discusses the question, Why should a person do what is most *likely* to work?

Suppose I am in a situation in which I have to do X or Y and the probability of success is very high if I do X and very low if I do Y. We can put Peirce's question this way: Why should I do X? Why is the fact that X will probably succeed a *reason* to do it?

The importance of Peirce's puzzle

Many philosophers would say that the reason one should be guided by the probabilities is that the *frequency of successes* one will enjoy will be higher if one does so. Observe that the case is not one in which the probabilities themselves are at all uncertain; we are supposed to know the probabilities, and so the problem of induction, that is, the problem of ascertaining the probabilities, is not the issue here. The issue is that we know the probability of success is high if one does X, low if one does Y, and the question is why should we do X? Observe also, that the given knowledge is of precisely the type

[3] This section of the present paper coincides with the conclusion of my Carus Lectures, *The Many Faces of Realism*, Open Court, 1987.

that is supposed to "justify" the hypothetical imperative "Do *X* if you want success".

It is at this point in the argument that Peirce's genius shows itself. Suppose that I am an old man, or that for some other reason I don't believe I have many years of life ahead of me. What do beliefs about what my success-frequency would be if I *were* to live a long time and be involved in a great many of these situations have to do with what I should do in this *one* situation? In fact, Peirce considers a situation in which the choice is between "eternal felicity" and "everlasting woe". By the very nature of this situation, there isn't going to *be* any *further* "gambling situation" which the rational agent will have to deal with. Specifically, Peirce's thought example is this:[4] one has to choose between two arrangements. Each arrangement is probabilistic; under each arrangement, one will select a card from a well shuffled pack with 25 cards in it, one of which is specially designated. The outcome depends in both cases on whether or not one draws the specially designated card. Under arrangement *A*, one gets everlasting woe if one draws the designated card and eternal felicity if one draws any other card, so that one's chances of eternal felicity are twenty four to one; while under the second arrangement it is the other way around —one gets eternal felicity if one draws the designated card and everlasting woe if one draws any other card, so that one's chances of everlasting woe are twenty four to one. (Those for whom the notion of immortality is troubling can substitute "an easy death" and "a hard death" for eternal felicity and everlasting woe, respectively.) We *all* believe that a rational person would choose arrangement *A*. Peirce's question is *Why should he*?

Reichenbach held that probability statements about the single case are simply a *fictitious transfer* of rela-

[4] Peirce discusses this example in "The Doctrine of Chances," p. 69; reprinted in *Chance, Love and Logic*, Morris R. Cohen (ed.) New York: Hartcourt, Brace, 1923.

tive frequencies in the long run,[5] or of knowledge of relative frequencies in the long run. Notice that this is yet another example of the use of the notion of a Projection; Reichenbach was saying that the very statement that Jones will have only one chance in twenty five of eternal felicity *this one time* under arrangement *A* is a "projection". There is no fact about the single unrepeatable situation which is *The fact that choice* A *gives Jones twenty four chances out of twenty five of eternal felicity.* (Recently, Stephen Leeds has written a stimulating paper[6] arguing that the whole notion of probability is a Projection.)

Pierce's problem comes out very clearly if we take the view that probability just is relative frequency in the long run. The person in the situation knows a fact which is *utterly irrelevant* to what he should do. He knows that *if there were* a series of situations like this one, then he *would have* eternal felicity twenty four times out of every twenty five if he were to choose arrangement *A* each time. But a person can have eternal felicity or everlasting woe only once! His problem is not how to achieve eternal felicity twenty four times *out of every twenty five*; his problem is to obtain a eternal felicity *this* time. Why should he pick arrangement *A*?

The only answer we can give is that it is more probable that he will have eternal felicity under arrangement *A*. But the question was, remember, Why should one *expect* what is *probable*? If you say that you should expect what is *probable* because it is *likely* to happen this time, you're not answering the question, you're just, as it were, repeating the advice: *Expect what is probable.* If you say, "Well, it's *reasonable* to expect what is probable," well —in this situation, isn't "reasonable" just a synonym for "probable," in the Keynes-Carnap sense of "logical

[5] Reichenbach discusses the single case in *The Theory of Probability,* 372ff.

[6] Leeds' paper was read at the Chicago meeting of the Philosophy of Science Association in November, 1984. It will appear in the proceedings of that meeting.

probability"? Isn't "It's reasonable to expect what is probable to happen," just another way of saying "It's *probable* (in the logical sense of probability) that what will probably happen (in the frequency sense of probability) will happen in any individual case (unless we know of some respect in which the individual case is atypical)?

We are forced back, then, to the view that a reasonable person adjusts his expectations to the *logical* probability; and this time, any beliefs we may have about how this will lead us to fare in the long run are seen to be irrelevant to the problem. That there is such a thing as the "logical probability," that it corresponds to the frequency in a long series (if there *were* a long series), and that a reasonable person adjusts his beliefs to *it* become just Ultimate Logical (*read:* metaphysical) Facts.

Peirce's own solution to this problem is one of the sources of inspiration for the views of Apel and Habermas that I mentioned in the last lecture. According to Peirce, one can *only* be rational if one *identifies* himself psychologically with a whole on going —in fact, a potentially infinite— community of investigators. It is only because I care about what *might* happen to people in similar situations that I do what has the best *chance* in my own situation. My belief that *I* in this one *unrepeatable* situation am somehow more likely to die easily than by torture is fundamentally, then, just what Reichenbach said it was, a fictitious transfer, on Peirce's view. What is true, and not fiction or projection, however, is that my fellows, the members of the community with which I identify, will have eternal felicity twenty four times out of twenty five if they follow this strategy; or more generally, even if this one particular situation is never repeated, that if in all the various *uncorrelated* cases of this kind or any other kind that they find themselves in they always follow the probabilities, then in the long run they will experience more successes and fewer losses.

But can it really be that the reason I would choose

arrangement *A* is that I am *altruistic*? Maybe I am, but isn't it obvious that I would choose arrangement *A* first and foremost because it would avoid everlasting woe *in my own case*? Peirce's argument is that I ought to choose arrangement *A* for what one might describe as "Rule Utilitarian" reasons: in choosing this arrangemen I am supporting, and helping to perpetuate, a rule which will benefit mankind (or the community of rational investigators) *in the long run*. Is this really what is in my mind when what I am facing is *torture* ("everlasting woe")? Frankly, it isn't. I cannot give a reason for doing what I would do in this case, if the only reasons allowed are in terms of "what will happen in the long run if". And this shows that even in the *means-end* kind of problem, I must fall back on intuitions that I am powerless to explain.

Today, many people[7] think that the only reason for being *reasonable* at all is that one will arrive at truth in theory and success in action *more often* if one is reasonable. Some people[8] have even proposed replacing the notion of a "reasonable" method by the notion of a reliable method: one that, as a matter of fact, leads to successfull outcomes with a high relative frequency. Notice that (if you agree with me in finding Peirce's own solution incredible) these approaches are helpless in the face of Peirce's problem. If my *only* reason for believing that I should be reasonable were my beliefs about what will happen *in the long run* if I act or believe reasonably, then I would have absolutely *no* reason (apart from the implausible reason of altruism) to think it better to be reasonable in an unrepeatable single case like the

[7] For a discussion of this kind of "epistemic utilitarianism" see Roderick Firth's presidential address, "Epistemic Merit, Intrinsic and Instrumental," in *Proceedings and Addresses of the American Philosopohical Society*, Sept. 1981, vol. 55, Number 1, 5-23.

[8] For a sophisticated version of this view see Alvin I. Goldman, "What is Justified Belief," in George Pappas (ed.), *Justification and Knowledge*, Boston, Dordrecht, London, 1979.

one described. In fact, as I came close to the end of my life, and found myself unable to make many more "bets," then my reasons for doing what is reasonable or expecting what is reasonable should diminish very sharply, on this view. The fact is that we have an *underived,* a *primitive* obligation of some kind to be reasonable, not a "moral obligation" or an "ethical obligation," to be sure, but nevertheless a very real obligation to be reasonable, which —contrary to Peirce— is *not* reducible to my expectations about the long run and my interest in the welfare of others or in my own welfare at other times. I *also* believe that it will work better in the long run for people to be reasonable, certainly; but when the question is *Why do you expect that, in this* unrepeatable *case, what is extremely likely to happen will happen?, here* I have to say with Wittgenstein, "This is where my spade is turned. This is what I do, this is what I say."

My reason for discussing this today, when the more usual question is what to do about the "bottomless pit" phenomenon in ethics, the lack of a Foundation in ethics, is that in the case just described —a case which has to do with reasonableness about "means and ends," rather than with ethics— my epistemic situation is exactly the same. I do think, and I think it warranted to think, that "acting on the probabilities" is the only rational thing to do, and that one ought to do the rational thing even in unrepeatable situations. In the ethical case, I do think, and I think it warranted to think, that a person who has a sense of human brotherhood is better than a person who lacks a sense of human brotherhood. A person who is capable of thinking for himself about how to live is better than a person who has lost or never developed the capacity to think for himself about how to live; but, wether the question be about single case probability or about ethics, I don't *know how I know* these things. These are cases in which I find that I have to say,

"I have reached bedrock and this is where my spade is turned."[9]

Recognizing that there are certain places where one's spade is turned; recognizing, with Wittgenstein, that there are places where our explanations run out, isn't saying that any particular place is *permanently* fated to be "bedrock," or that any prticular belief is forever immune from criticism. This is where my spade is turned *now.* This is where my justifications and explanations stop *now.* To recognize that a loyal human being is better than a disloyal human being, that a person capable of *philia* is better than a person incapable of *philia,* that a person capable of a sense of community, of citizenship in a *polis,* is better than a person who is incapable of a sense of community or of citizenship in a *polis*, and so forth, is not to say that any one of these values or any one of the moral pictures which may lie behind and organize these values is final, in the sense, of being exclusively or exhaustively correct. Our moral images are in a process of development and reform. But it is to say that at each stage in that development and reform, there will be places, many places, at which we have to say, "This is where my spade is turned."

None of this goes against the idea that rational criticism of a moral vision is possible. A moral vision may contradict, for example, what we know or think it rational to believe on other grounds, be they logical, metaphysical, or empirical. But we cannot any longer hope that these kinds of criticism will leave just *one* moral vision intact. Ultimately, there is still a point at which one has to say, "This is where my spade is turned."

[9] *Philosophical Investigations,* sec. 217. That Wittgenstein here uses the first person —where *my* spade is turned— is very important; yet many interpreters try to see his philosophy as one of simple deference to some "form of life" determined by a community. On this, see also Stanley Cavell's fine discussion in *The Claim of Reason,* esp. Part One, Chapter V: "The Natural and The Conventional."

RESUMEN

Se argumenta que es erróneo suponer que las dificultades, al explicar la racionalidad de enunciados en ética y en ciencia, no son problemas de la misma naturaleza. El autor sostiene que la racionalidad es algo tan difícil de explicar en el campo científico como en el ético, en especial, en las "instrumentaciones" tanto como en las "valoraciones". Se ocupa de cuestionar la teoría de la decisión desde dos perspectivas: por una parte, el aparato formal elaborado por von Neumann y, por la otra, las nociones intuitivas que cualquier agente racional desearía defender, *i.e.* "autonomía", "indiferencia", "incompatibilidad", que han sido canceladas u oscurecidas por la teoría clásica de la decisión. Putnam concluye que pese a no poseer ninguna tesis positiva más clara al respecto y, pese a su crítica de la teoría de la decisión, reconoce que es racional actuar guiados por las probabilidades y que hay cuestiones en las cuales la filosofía parece tocar fondo y en donde sus explicaciones y justificaciones se agotan, pero que esto no es decir que la crítica racional de una visión moral no sea posible ni que todos los problemas hayan sido cancelados.

Sobre el aparato formal, se ocupa de los axiomas de "preferencia racional" para un agente racional ideal, que implican que todas las elecciones —incluyendo las que intuitivamente son elecciones alternativas "incomparables"— pueden ordenarse de tal forma que, para cualesquiera dos productos, o bien (a) no son iguales (se prefiere uno sobre el otro), o bien (b) son iguales (resultan "indiferentes" al sujeto). La discusión gira en torno de (b), en virtud de que a menudo uno desea intuitivamente diferenciar entre alternativas "incomparables" y alternativas "indiferentes". Sostiene que soslayar esta distinción permite a la teoría de la decisión desechar una característica importante en la decisión racional, a saber: la autonomía.

De acuerdo con Putnam esta distinción intuitiva se cancela no sólo por la aplicación que hace von Neumann del axioma de preferencia, según el cual (usando "xPy" para simbolizar "x es preferible a y"):

$$(xPy \ \& \ yPz) \text{—} xPz$$

es decir, se afirma el supuesto *no* objetable para un agente ideal racional de que la preferencia es transitiva; *sino* por asumir que el complemento de tal axioma *también es transitivo* (esto es: si x no es preferible a y y y no es preferible a z, entonces x no es preferible a z):

$$(\text{-}xPy \ \& \ \text{-}yPz) \text{—} \text{-}xPz$$

Señala Putnam que el teórico de la decisión sostiene entonces algo muy fuerte y que lo hace sobre la base de que en caso de violar tal axioma, el agente sería visto como irracional en virtud de que estaría obligado a admitir un argumento "*Dutch book*", esto es, estaría obligado a admitir un sistema de apuestas de acuerdo con el cual el sujeto no podría obtener ninguna apuesta favorable. Para enfrentar este razonamiento, Putnam presenta un contraejemplo al argumento, tomado de Peirce, en el cual se muestra que el agente estaría obligado a enfrentarse a un "*Dutch book*" sólo porque el teórico de la decisión ha supuesto que el agente racional ideal ha hecho ya todas sus apuestas existenciales, lo cual es claramente absurdo. Por otra parte, muestra, con el mismo ejemplo de Peirce, que no es viable hacer sobre la base de probabilidades, una elección racional sobre una única alternativa irrepetible, en virtud de que tales probabilidades son una mera "proyección". Finalmente, argumenta que no debe considerarse que tal "proyección" está sujeta a razones de "reglas utilitaristas" en ética, puesto que nuevamente dejan fuera un elemento importante: mi decisión es mía, en virtud de lo que a *mí* y sólo a *mí* pueda sucederme.

[Lourdes Valdivia]