PATRICK SUPPES

SOME FORMAL MODELS OF GRADING PRINCIPLES*

I. INTRODUCTION

The present paper offers an analysis of grading principles from the viewpoint of statistical decision theory and game theory. The mistaken notion is widely held that the plain man is really clear about practical ethical and moral issues and that philosophers need only tidy up certain wayward corners of the subject.[1] Personally I find difficult the problem of devising any general ethical rules of behavior for simple two-person games; the ethical complexities of progressive taxation, tariff barriers, or treatment of sexual psychopaths are beyond any exact conceptual analysis. That decisions are and must be made about these issues no more proves that their ethical aspects are completely understood than does the fact that the Romans built bridges prove that they had any quantitative grasp of the mechanical theory of stress.

It is pertinent to remark that the first model used in this paper is at the basis of much recent foundational work in statistics (see Blackwell and Girshick (1954) and Savage (1954)). The considerations in the last two sections are within the more general framework of the theory of games as developed by von Neumann and others. My particular concern is the embedding in this framework of a theory of two-person justice.

II. INDIVIDUAL DECISION MODEL

The structure of the first model to be considered is simple. We shall call an

* I am indebted to Richard Brandt, Donald Davidson and F. Studnicki for a number of useful and penetrating criticisms of a much earlier draft of this paper written in 1957 and circulated as a technical report in that year under the title, 'Two formal models for moral principles'.

[1] Kant's views are typical: "... in matters which concern all men without distinction, nature cannot be accused of any partial distribution of her gifts; and that with regard to essential interests of human nature, the highest philosophy can achieve no more than that guidance which nature has vouchsafed even to the meanest understanding" (1949a, p. 666).

Synthese **16** (1966) 284–306. © *D. Reidel Publishing Co., Dordrecht-Holland*

ordered triple $\mathscr{S} = <S, C, D>$ an *individual decision* situation when S and C are sets and D is a set of functions mapping S into C. The intended interpretation is:

S = set of states of nature,
C = set of consequences,
D = set of decisions or actions.

Since the terms 'states of nature', 'consequences', 'decisions' and 'actions' are used here in a somewhat special manner, an example may help to make clearer their intended meaning.

EXAMPLE 1: Suppose I come home and find a bottle of ink spilt on the rug, and also suppose I know immediately that it could have been spilt either by my four-year old daughter or by my cat. These two possibilities correspond to the two states of nature. I can take one of two actions, let us say: spank the child or do not spank the child. And the possible consequences are four in number, as illustrated in Table I. The rows correspond to the two states of nature, the columns to the two actions, and the entries in the table to possible consequences.

TABLE I

actions / states of nature	a_1 – spank the child	a_2 – do not spank the child
s_1 – child spilt the ink	c_1 – ink spilt by child and child spanked	c_2 – ink spilt by child and child not spanked
s_2 – cat spilt the ink	c_3 – ink spilt by cat and child spanked	c_4 – ink spilt by cat and child not spanked

Since the term 'states of nature' is not much used in philosophy there should be little objection to its special use here; the term 'action' is used in a way that is consonant with at least one of its major uses in ordinary contexts. But my use of 'consequence' is probably at variance with its primary use in the writings of moral philosophers. The consequence c_1 above, for instance, ink spilt by child and child spanked, would be regarded by many as the bare beginning of consequences. It is to avoid

exactly the vagueness of the consequences flowing from c_1, c_2, c_3 or c_4, that I have adopted the restricted use. The longer term 'immediate consequence' could be used. Yet in ordinary usage there is much to defend the use adopted here. When a quarterback throws an intercepted pass in the last two minutes of play it might be appropriate to remark "The consequence of that is obvious. We lose the game." It would seem pedantic to insist on saying "The *immediate* consequence of that is obvious. We lose the game." And it would be a classroom gambit to object that the use of the definite article is wrong, because the action could have other important consequences for the quarterback: he quarrels with his girl that night, the coach decides not to start him in the next game.

Apart from any questions of ordinary usage there is a technical device which may be used to meet the difficulty that it is almost always impossible to characterize the full set of consequences which may flow from an action. Given an individual decision situation $\langle S, C, D\rangle$, let C' be the set of all consequences which result from some state of nature in S and some decision in D. Then C is a *partition* of C', that is C is a family of non-empty pairwise disjoint sets whose union is C'. In this analysis, each c_i in our example is a set of consequences. It is practically impossible to say exactly what the members of c_1, say, are, but in rough terms they are the possible consequences, proximate and remote, which would wholly or in part result from the immediate consequence of the ink's being spilt by the child and the child's being spanked.[2]

The still more complicated question of what kind of language is appropriate for describing either consequences or states of nature cannot be examined here. Certainly in most situations it is difficult to avoid evaluative or normative terms, but the use of non-factual language does not directly disturb or vitiate the analysis given here.

One of the basic problems of statistical decision theory is to introduce a preference ordering on the set of decisions of an individual decision situation and to consider what postulates the preference ordering of a reasonable man should satisfy. (For such an analysis see Savage (1954) or Suppes (1956).) The notion of reasonableness or rationality used here is an informal, intuitive one, and its application in defense of any particular

[2] I emphasize that consequences are to be construed broadly here. Causal as well as logical relationships are relevant, but an exact discussion of the significance of causal concepts in the present context would require too lengthy a digression to be appropriate.

postulate consists of analyzing particular examples. The problem is presumed solved if reasonable postulates can be found which are strong enough to guarantee the existence of a (subjective) probability measure on the states of nature and a utility function on the set of consequences such that one decision is to be preferred to another if and only if the expected utility of the first decision with respect to the probability measure is greater than that of the second. Once such a probability measure and utility function are constructed no further principles of action are needed. The sole maxim to be followed by the rational man is: maximize expected utility.

Historically the idea of maximizing utility is closely connected with the hedonistic ideas of Bentham, Mill, Sidgwick, and their followers. However, it is an unequivocal mistake to think that the maxim: maximize expected utility, in any respect involves a commitment to hedonism. As I hope to make clear in the sequel if the utility function on consequences were guided by an ethic of duty rather than pleasure, it would still be good advice to maximize expected utility. In this case a calculus of duty would replace a calculus of pleasure. To my mind the most important aspect of the hedonistic tradition in ethics has been the clear recognition that *some* principle of calculation is required for rational action in the face of other than trivial situations. The main point of this paper is to defend a thesis as to how grading principles should enter into these calculations.

Before developing these ideas further I want to say something about a major criticism that is usually made of the general maximization viewpoint adopted here. To wit, as one philosopher scornfully put it to me, whoever heard of a man making such calculations prior to making any actual decision. Naturally this philosopher had in mind the "ordinary" man in "ordinary" situations like that of buying a pint of whiskey or selecting a new tobacco. One might as well reject a whole discipline such as the physical theory of the strength of materials by remarking that no carpenter computes the load capacity of a joist before sawing and nailing it. There are situations where elaborate calculations are made in order to maximize utility; the new disciplines of management science and operations research provide numerous examples.[3] Moreover, I maintain that in

[3] In this respect it seems unfortunate that in his inaugural lecture *Theory of Games as a Tool for the Moral Philosopher* Professor Braithwaite picked for detailed analysis an

many ordinary situations it is not the impossibility of detailed calculation that is relevant but rather the superfluity of it. For instance, in the simple situation schematized by Table I, if it is definitely known that the ink was spilt by the child and not the cat then to take appropriate action I need only order in preference two consequences: c_1 and c_2, according to my principles of childrearing. I need no numerical utility function. And this situation is characteristic: whenever uncertainty regarding the true state of nature is eliminated, the pertinence of a numerical utility function disappears, and the principle of maximizing expected utility assumes a very simple form: choose that action whose consequence is most preferred (for reasons of pleasure, duty, justice, or what have you).

III. DEFINITION OF GRADING PRINCIPLES

Traditionally in ethics, actions are said to be right or wrong, and consequences good or bad. If we carried over this distinction to individual decision situations then we would need moral principles of grading governing acts and value principles of grading arranging consequences in order of preference. But I am proposing here that the one controlling moral principle of action is the maxim: maximize expected utility. (Hereafter referred to as the M.E.U. maxim.) On this view it is a mistake to hold that grading principles aid us directly in distinguishing between the quality of acts. The function of grading principles is rather to aid the individual in constructing his preference relation on the set of consequences.[4] There is a simple reason why this position is not in conflict with most of the standard examples purporting to show how grading principles should regulate actions; namely, if the state of nature is known, there is an effective one-one correspondence between the set D of acts and the set C of consequences, and any relation on C defines a corresponding relation on D. This point is further amplified below.

To put it baldly then, I am claiming that the proper logical status of a grading principle in an individual decision situation is as a binary relation

example which would not in practice be subject to elaborate calculations. His painstakingly careful presentation would apply equally well to more realistic labor-management bargaining situations.

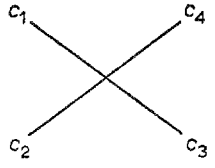
[4] However, second-order moral principles as ethical rules of behavior directly governing acts are introduced in the final section.

on the set C of consequences, in fact, an asymmetric, transitive relation on C, i.e., a strict partial ordering of C.

DEFINITION 1: *Let $\mathscr{S} = \langle S, C, D \rangle$ be an individual decision situation. Then a grading principle with respect to $\mathscr{S}$ is a strict partial ordering of C.*

I have insisted that a grading principle have at least the properties of a strict partial ordering, for otherwise it would scarcely be a guide to fixing the preference relation.

EXAMPLE 2: *A principle of childrearing.* Referring to Example 1, a tenable grading principle held by some modern parents is: never punish a child. This leads to the following strict partial ordering of C[5], which we may represent by a Hasse diagram:



It should be noted that this principle of childrearing is sufficient to determine action although the set of consequences is not completely ordered by it. For, whatever the true state of nature, the consequence of taking action a_2 is preferred to the consequence of taking action a_1, that is, c_2 is preferred to c_1 and c_4 is preferred to c_3.

As the following example drawn from welfare economics shows, most grading principles are not sufficient to determine action.

EXAMPLE 3: *Principle of unanimity.* Suppose that the decision situation consists of an arbitrary set S of states of nature, and C is a set of ordered n-tuples (n-dimensional vectors) representing the distribution of some desired commodity to a group of n individuals. Administrator A, a member of the group, is to decide in a just manner which distribution vector is to be used in allotting the quantities of the commodity. The grading principle of unanimity asserts that vector $x = \langle x_1, \ldots, x_n \rangle$ is to be preferred to vector $y = \langle y_1, \ldots, y_n \rangle$ if for every $i = 1, \ldots, n, x_i \geq y_i$ and for some $i, x_i > y_i$. This principle, also known as the principle of efficiency or Pareto optimality, is a very weak grading principle and surely any adminis-

[5] The intuitive idea behind a Hasse diagram is simple: if point x may be reached from point y by a continually ascending, not necessarily straight line, then xGy.

trator who did not satisfy it would be stoned out of office.[6] It is obvious that in general the principle of unanimity does not uniquely determine the optimal action even when there is only one state of nature.

More troublesome, at least from a psychological standpoint, is the decision situation in which two grading principles are in conflict. This state of affairs is reflected formally in our model by the fact that the union of two strict partial orderings is not always a strict partial ordering.

DEFINITION 2: *Let $\mathscr{S} = \langle S, C, D \rangle$ be an individual situation, and let G_1 and G_2 be grading principles with respect to $\mathscr{S}$. Then G_1 and G_2 are compatible if, and only if, $G_1 \bigcup G_2$ is a grading principle with respect to $\mathscr{S}$.*[7]

Two simple conditions with reasonable interpretations which will insure compatibility of grading principles are the following.

THEOREM 1: *If* (i) *G_1 is a subrelation of G_2 or G_2 is a subrelation of G_1, or* (ii) *if the fields of G_1 and G_2 are mutually exclusive, then G_1 and G_2 are compatible.*

The proof of this theorem is trivial. Some examples illustrating it may be drawn from welfare economics, where S and C are defined as in Example 3.

EXAMPLE 4: Let

G_1 = principle of unanimity,
G_2 = principle of gross aggregation,
G_3 = principle of social weights $a = \langle a_1, \ldots, a_n \rangle$,

where

$$xG_2y \quad \text{if, and only if,} \quad \sum_{i=1}^{n} x_i > \sum_{i=1}^{n} y_i$$

and

$$xG_3y \quad \text{if, and only if,} \quad \sum_{i=1}^{n} a_i x_i > \sum_{i=1}^{n} a_i y_i .$$

Principle G_1 has already been discussed; Principle G_2 says that one distribution x is to be preferred to another y if x results in a greater total quantity of the commodity for the social group; Principle G_3 corresponds

[6] The proof is immediate that the principle of unanimity yields a strict partial ordering of C.

[7] The symbol $\cup$ denotes the union of two sets. A binary relation is a set of ordered couples, whence we may speak of the union of two relations.

to the assignment of weights to each individual by Administrator A; presumably A would use some further principle of need or merit to aid in determining the weights.[8] As application of Theorem 1, we have that G_1 and G_2 are compatible, since G_1 is a subrelation of G_2, that is, if xG_1y then xG_2y. To see this, we observe that if xG_1y then

(1) $x_i \geq y_i$ for all i
$x_i > y_i$ for some i,

whence

$$\sum x_i > \sum y_i.$$

Moreover, if each individual is given a strictly positive weight, that is, $a_i > 0$ for all i, then G_1 is a subrelation of G_3, and hence compatible with it. The reasoning is obvious. From (1) and the hypothesis that $a_i > 0$ we have:

$a_i x_i \geq a_i y_i$ for all i
$a_i x_i > a_i y_i$ for some i,

whence by addition of inequalities

$$\sum a_i x_i > \sum a_i y_i.$$

On the other hand, when C has any abundance of different distribution vectors, G_2 and G_3 are incompatible. For instance, let $n=3$ and

$$x = <1, 2, 4>$$
$$y = <4, 1, 1>$$
$$a = <2, \tfrac{1}{2}, \tfrac{1}{4}>.$$

Then

$$xG_2y$$

since

$$\sum x_i = 7 \text{ and } \sum y_i = 6,$$

but

$$yG_3x$$

[8] In the literature of socialist economics, Administrator A is often the Central Planning Board, but a bureaucratic assignment of weights is not essential to the economic theory of the welfare state (cf. Lange and Taylor, 1938).

since

$$\sum a_i x_i = 4 \text{ and } \sum a_i y_i = 8\tfrac{3}{4}.$$

Examples which satisfy (ii) of Theorem 1 are easy to construct but will not be considered here. The intuitive idea of (ii) is the truism that grading principles concerned with entirely different spheres of activity are compatible.

The use of the word 'activity' in the last sentence underlines the difficulty of not speaking of grading principles as referring to acts or decisions rather than consequences. Before turning to social decision situations in the next section, something more needs to be said about the status here advocated for grading principles. One natural tendency is to formulate grading principles in the imperative mood so as to command the execution of certain acts. But Hare (1952, Part III) has cogently argued it is more appropriate to use the indicative mood and the auxiliary verb 'ought' to obtain the proper sort of universal formulation. The one further emendation required here is to add the infinitive 'to prefer' after 'ought'. Thus, we go from the imperative:

'Honor thy father'

to:

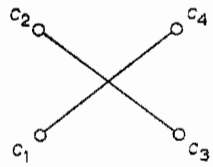
'Everyone ought to honor his father',

and on to:

(1) 'Everyone ought to prefer to honor his father.'

I maintain that ordinary usage addresses moral principles of grading directly to acts because the problem of acting without knowing the true state of nature is ignored. This point is important enough to be amplified by referring again to Example 1. Consider the moral imperative 'Punish the guilty and defend the innocent'. Suppose this is the only moral imperative guiding my choice of action a_1 or a_2 in Example 1. It seems patently obvious that without knowing the true state of nature I can make no direct application of the imperative to choose between a_1 and a_2. If s_1 is the true state of nature I should choose a_1, but if s_2 is the true state, I should choose a_2. To be sure, I could first sum up the factual evidence for s_1 and s_2, decide which is more likely, assume the more likely state is nearly certain to be the true state, and then take the appropriate action.

But this is surely a crude way to proceed and is wholly inadequate in more complicated situations; for instance, suppose there were three states of nature to each of which I assigned a subjective probability of 1/3. On the other hand, the imperative may be applied directly to constrain my preference relation on the set $\{c_1, c_2, c_3, c_4\}$ of consequences. The Hasse diagram of the resulting strict partial ordering is obviously:



which may be compared with the diagram for Example 2. When applied directly to consequences, application of the imperative need not be confounded with the difficult and distinct problem of weighing factual evidence regarding the true state of nature.

The particular homily about honoring fathers illustrates another point: it and all principles of a similar form lead to a simple and crude partial ordering of consequences, namely, consequences are divided into two classes and all members of one are preferred to all members of the other. As Examples 3 and 4 emphasize such principles are not of much help in making a rational decision in a complicated situation like that generated by a labor-management dispute or the problem of pricing policy in a semi-controlled economy.

IV. SOCIAL DECISION MODEL

Against the analysis of previous sections may be brought the charge that the individual decision model unduly and unrealistically isolates the behavior of one man from another. In the remainder of this paper social situations shall be considered. For reasons of technical simplicity the discussion shall be restricted to two persons, although most of the concepts introduced readily generalize to n persons.

The structure of the basic model is still relatively simple. We shall call an ordered sextuple $\mathscr{S} = \langle S, C_1, C_2, D_1, D_2, f \rangle$ a *two-person decision situation* when S, C_1, C_2, D_1, D_2 are sets and f is a function mapping the

Cartesian product $S \times D_1 \times D_2$ into $C_1 \times C_2$. The intended interpretation is:

S = set of states of nature,
C_1 = set of consequences for person I,
C_2 = set of consequences for person II,
D_1 = set of decisions or acts available to I,
D_2 = set of decisions or acts available to II,
f = social decision function.

Some examples will be given in the next section in connection with the theory of two-person justice.

The definition of grading principles is an obvious generalization of the one already given for the individual case.

DEFINITION 3: *Let* $\mathscr{S} = \langle S, C_1, C_2, D_1, D_2, f \rangle$ *be a two-person decision situation. Then a grading principle with respect to* $\mathscr{S}$ *is a strict partial ordering of the Cartesian product* $C_1 \times C_2$.

This definition does not require that in applying a grading principle person I need consider consequences to person II, but does make possible such a consideration. We could in fact use Definition 3 as a basis for defining a wholly egocentric person, namely a person, say I, whose grading principles and preference relations in all two-person situations are orderings uniquely determined by elements of C_1 (the character of C_2 being never considered).

The same arguments given previously apply to the requirement that ordinary grading principles in two-person situations be partial orderings on consequences and not on acts. On the other hand, the arguments for a rigid adherence to the M.E.U. maxim are not so persuasive, since other rules of behavior like minimax or minimax regret can be strongly defended for two-person situations. But these matters will not be gone into here; for our purposes adoption of any of these alternative rules requires admission only that a utility or value function on $C_1 \times C_2$ is needed for both persons I and II. We want to investigate how a formal principle of justice may be introduced which will put non-trivial constraints on the utility function. Moreover, it will be of interest to investigate the adequacy of a justice maxim compared to the M.E.U. or minimax kind of maxim, as an over-all rule of behavior.

In concentrating attention on justice no claim is intended that it is the most significant grading principle for social situations, nor even that

the definitions given here provide more than the merest beginning of a formal theory of justice.

To begin with we need the notion of a preference relation on $C_1 \bigcup C_2$, that is, on the set of consequences to both I and II. The idea is that one consequence in $C_1 \times C_2$ will be deemed more just or fair than another relative to a preference ranking of all consequences together. How in fact would a person make such a ranking? Presumably by treating himself and the other person on an "equal" basis. A suggestion as to how this idea of equality or symmetry may be formalized will be given the following definition:

DEFINITION 4: *A system* $\mathcal{S} = \langle S, C_1, C_2, D_1, D_2, f, R_1, R_2 \rangle$ *is a two-person decision situation with preference rankings if, and only if,* $\langle S, C_1, C_2, D_1, D_2, f \rangle$ *is a two-person decision situation, and* R_1 *and* R_2 *are weak orderings of* $C_1 \bigcup C_2$. (A weak ordering is a relation which is transitive and strongly connected.)

The intended interpretation is that R_1 is the preference ranking of person I and R_2 that of II. Formally we might say that a person's preference ranking R of $C_1 \bigcup C_2$ is equitable or symmetric if it remains unchanged when the two persons change positions in the decision situation. Difficulties of making this suggestion precise will not be pursued here, but it would seem best to do it in terms of a specific game, or at least gamelike, structure, with the exchange being defined in terms of becoming a different player in the game, not, by all means, in terms of the personal attributes of the players somehow being exchanged. It is intended that in constructing R_1, say, person I will say to himself, it is better that II have x in C_2 than that I have y in C_1 whence xR_1y and not yR_1x, etc. For example, a man should judge it better that his neighbor of equal economic status receive a thousand dollars than that he himself should receive fifty dollars. Unfortunately, I see no way of characterizing in an adequate formal manner the intuitive notion of *better than* used in this example. But it would be a mistake to consider this situation peculiar to moral philosophy. The notion of preference or *better than* has a status in formal moral philosophy very similar to that of the notion of force in mechanics. It is not a problem of mechanics proper to classify forces according to their physical origin.[9]

[9] These remarks are admittedly Kantian in flavor. Cf., "And just as nothing follows from the primary formal principles of our judgments of truth except when primary

We now define the notion of *more just than* (abbreviated by J) relative to each person's preference ranking.

DEFINITION 5: *If* $x_1, y_1 \in C_1$ *and* $x_2, y_2 \in C_2$ *and* $x = \langle x_1, x_2 \rangle$ *and* $y = \langle y_1, y_2 \rangle$ *then for* $i = 1, 2$, xJ_iy *if, and only if, either* (i) $x_1 R_i y_1$ *and* $x_2 R_i y_2$ *and not* ($y_1 R_i x_1$ *and* $y_2 R_i x_2$), *or* (ii) $x_1 R_i y_2$ *and* $x_2 R_i y_1$ *and not* ($y_2 R_i x_1$ *and* $y_1 R_i x_2$).

This definition is simpler than it may appear at first glance. It is framed so as to make J_i (for $i = 1, 2$) a strict partial ordering of the Cartesian product $C_1 \times C_2$, and yet permits the comparison of elements of C_1 with C_2. The two "not" clauses in the definition guarantee that J_i is asymmetric.

Examples of J_i are at the beginning of the next section. We conclude this section with the theorem:

THEOREM 2: *Both* J_1 *and* J_2 *are grading principles with respect to* $\mathscr{S}$.

PROOF: For $i = 1, 2$, to prove that J_i is asymmetric, suppose by way of contradiction that for some $x = \langle x_1, x_2 \rangle$ and $y = \langle y_1, y_2 \rangle$ in $C_1 \times C_2$ that

$$xJ_iy \text{ and } yJ_ix.$$

From xJ_iy it follows from the definition that (dropping subscript i on R for brevity) x_1Ry_1 or x_1Ry_2, and similarly from yJ_ix it follows that y_1Rx_1 or y_1Rx_2. We thus have four cases to consider:

Case 1: x_1Ry_1 and y_1Rx_1. Case 3: x_1Ry_2 and y_1Rx_1.
Case 2: x_1Ry_1 and y_1Rx_2. Case 4: x_1Ry_2 and y_1Rx_2.

Since the proofs for all cases are similar, we shall look only at Case 2 in detail. From the hypothesis of this case, we have from (i) of the definition:

(1) x_1Ry_1,
(2) x_2Ry_2,
(3) not y_1Rx_1 or not y_2Rx_2,

and from (ii):

(4) y_1Rx_2
(5) y_2Rx_1
(6) not x_2Ry_1 or not x_1Ry_2.

---

material grounds are given, so also no particular definite obligation follows from these ... rules except when indemonstrable material principles of practical knowledge are connected with them" (Kant, 1949b, pp. 283–284).

From (1), (4) and the transitivity of R we infer:

(7) $x_1 R x_2$

and from (7) and (2):

(8) $x_1 R y_2$.

From (2) and (5) (and transitivity of R):

(9) $x_2 R x_1$,

and from (9) and (1):

(10) $x_2 R y_1$,

but (8) and (10) contradict (6).

To prove now that J_i is transitive, we assume

$$x J_i y \text{ and } y J_i z,$$

which leads to four cases also. Again we shall consider only one typical case:

(11) $x_1 R y_2$ and $y_1 R z_2$.

From (11) and (ii) of the definition, we have:

(12) $x_2 R y_1$
(13) $y_2 R z_1$
(14) not $y_2 R x_1$ or not $y_1 R x_2$
(15) not $z_2 R y_1$ or not $z_1 R y_2$.

From (11), (13) and transitivity of R, we get:

(16) $x_1 R z_1$,

and similarly from (11) and (12):

(17) $x_2 R z_2$.

It remains to show that not $z_1 R x_1$ or not $z_2 R x_2$. Suppose by way of contradiction that

(18) $z_1 R x_1$ and $z_2 R x_2$.

From (18) and (11), we have:

(19) $z_1 R y_2$,

and from (18) and (12)

(20) $\quad z_2 R y_1,$

but (19) and (20) contradict (15), which completes our proof, since for J_i to be a grading principle with respect to $\mathscr{S}$, it is required by definition that J_i be asymmetric and transitive in $C_1 \times C_2$.

V. POINTS OF JUSTICE AND THE PRISONER'S DILEMMA

It will be instructive to apply the ideas introduced in the last section to a simple but conceptually troublesome example of a two-person, non-zero-sum, non-cooperative game known as the prisoner's dilemma.[10] We quote the description from Chapter 5 of Luce and Raiffa (1957):

Two suspects are taken into custody and separated. The district attorney is certain that they are guilty of a specific crime but he does not have adequate evidence to convict them at a trial. He points out to each prisoner that each has two alternatives: to confess to the crime the police are sure they have done, or not to confess. If they both do not confess, then the district attorney states he will book them on some very minor trumped-up charge such as vagrancy and they will both receive minor punishment; if they both confess they will be prosecuted, but he will recommend less than the most severe sentence; but if one confesses and the other does not, then the confessor will go free while the latter will get "the book" slapped at him.

Let $\quad n$ = no conviction on any charge,
v = vagrancy conviction,
r = reduced conviction (less than maximum),
m = maximum conviction.

Then the game may be represented by:

I \ II	confess	not confess
confess	$< r, \ r >$	$< n, m >$
not confess	$< m, n >$	$< v, \ v >$

[10] A game is non-cooperative when no precommunication or bargaining between the players is permitted. The prisoner's dilemma is attributed to A. W. Tucker.

where a pair like $\langle n, m \rangle$ is interpreted so that the first member n is the outcome to person I and the second member m the outcome to person II. We have not distinguished n_{I} and n_{II}, m_{I} and m_{II}, etc. These consequences are treated the same for each player. Keeping this in mind, the complete two-person decision situation with preference rankings may be identified, provided we introduce the one obvious preference ranking on the set of consequences:

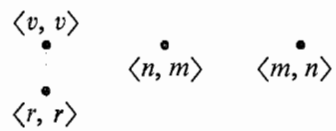
S = one element set (trivial here),
$C_1 = C_2 = \{m, n, r, v\}$[11],
$D_1 = D_2 = \{\text{confess, not confess}\}$,
f = function defined by above game matrix,
$R_1 = R_2$ = weak ordering arising from linear ordering n, v, r, m, with n most preferred.

Clearly here

$$J_1 = J_2,$$

and the ordering *more just than* of $C_1 \times C_2$ may be represented by the following Hasse diagram (see next page), where two elements of $C_1 \times C_2$ standing at the same point in the diagram are not comparable under J_i.[12] Of course, only the four elements in the game matrix are of direct concern in discussing the prisoner's dilemma. The ordering induced by J_i on them may be represented by:

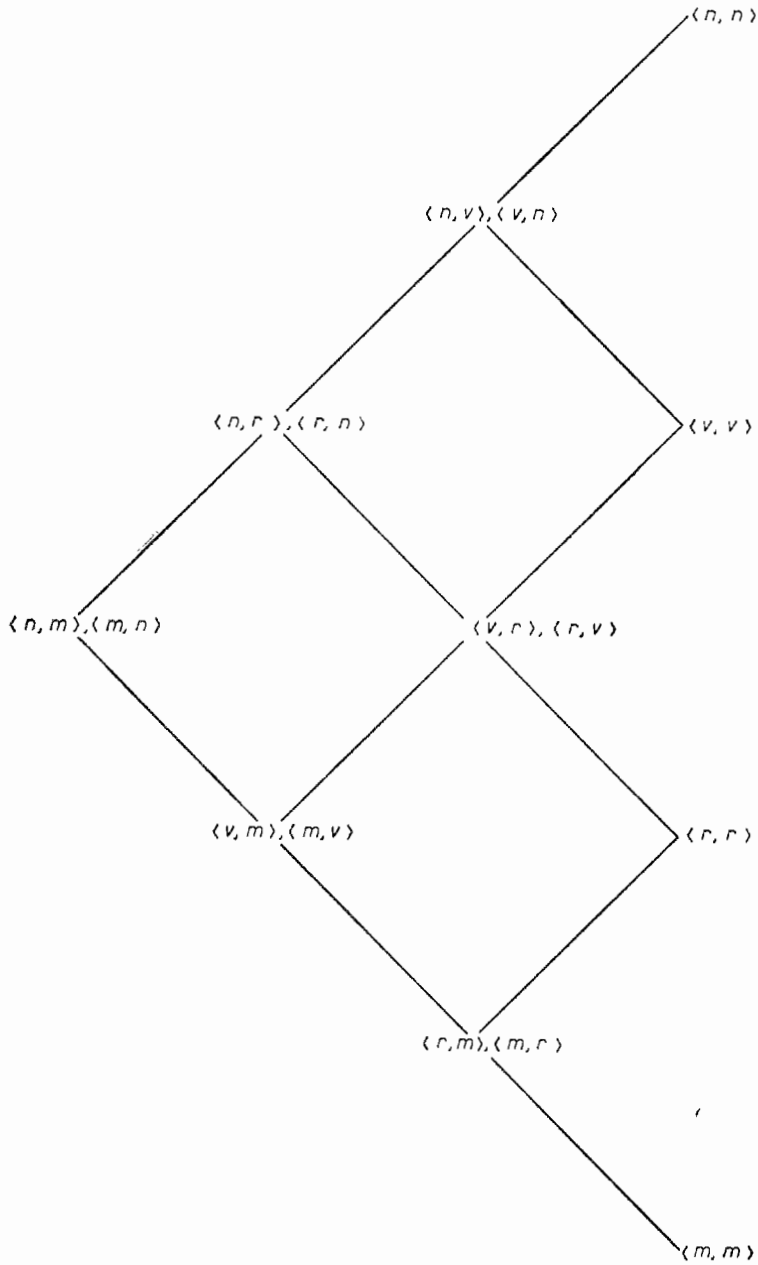
(1)



The important thing is that $\langle n, m \rangle$ is not related by J_i to any of the other three elements, nor is $\langle m, n \rangle$.

The weak relation expressed by (1) would not seem to be of much help in guiding the choice of an action or strategy by either prisoner. As a direct constraint on the utility function of either it scarcely imposes any structure. Before attempting to show that considerably more can be

[11] The identification of C_1 and C_2 merely simplifies the presentation and is not essential.
[12] Under the equivalence relation which "identifies" elements like $< r, m >$ and $< m, r >$, $C_1 \times C_2$ is a lattice with respect to J_i, but this fact is of no significance here.


⟨n, n⟩
⟨n, v⟩, ⟨v, n⟩
⟨n, r⟩, ⟨r, n⟩
⟨v, v⟩
⟨n, m⟩, ⟨m, n⟩
⟨v, r⟩, ⟨r, v⟩
⟨v, m⟩, ⟨m, v⟩
⟨r, r⟩
⟨r, m⟩, ⟨m, r⟩
⟨m, m⟩

obtained from (1) by introducing the concept of a *point of justice*, it will be useful briefly to review the game-theoretic solution of the prisoner's dilemma.

Two concepts of optimality for two-person, non-zero-sum, non-cooperative games yield the conclusion that both prisoners should choose the strategy of confessing, which leads to the outcome or consequence $\langle r, r \rangle$. One concept arises from the highly appealing *sure-thing* principle. A strategy or decision satisfies the sure-thing principle if no matter what your opponent does you are at least as well off, and possibly better off, with this strategy in comparison to any other available to you. Thus if person I adheres to the sure-thing principle he should confess, for if II confesses I gets r rather than m, and if II does not confess I gets n rather than v; whence for every choice of II, I is better off confessing. A similar situation obtains for II.

In many games no strategy satisfies the sure-thing principle. But every finite game of the class being discussed does have at least one *equilibrium point*, the second concept of optimality (introduced by John Nash (1950), (1951)). Roughly speaking, an equilibrium point is a set of strategies, one for each player, with the property that these strategies provide a way of playing the game such that if all the players but one follow their given strategies, the remaining player cannot do better by following any strategy other than the one belonging to the equilibrium point. As is easily verified, the unique equilibrium point for the prisoner's dilemma is the pair of confession strategies, the same result obtained by application of the sure-thing principle.

In spite of the weight of these optimality principles there are several unsatisfactory aspects of the recommended solution. If both prisoners completely trust each other it seems more reasonable for both of them to adopt the strategy of not confessing. Moreover, the act of confessing might from a moral standpoint be distasteful. The various game-theoretical principles of behavior like the two just discussed are aimed at satisfying intuitive ideas of prudential rather than moral behavior – the notion of prudence being that of acting in one's own best interest without direct concern for others. The point of the remainder of this paper is to contrast moral and prudential behavior, with special reference to the prisoner's dilemma.

In Section III I have argued that grading principles should be addressed

to consequences rather than decisions or acts. I now want to suggest that a (first-order) grading principle concerned with consequences may lead to a (second-order) moral principle which is a direct guide to action. Such second-order moral principles may be termed *ethical rules of behavior*, in contrast to game-theoretical *prudential* principles of behavior. I shall use the justice relation on consequences to formulate one such ethical rule of behavior. First we define a (J_i) *admissible element* as an element of $C_1 \times C_2$ which is not dominated under the relation J_i by any other element. In diagram (1) elements $\langle v, v \rangle$, $\langle n, m \rangle$ and $\langle m, n \rangle$ are (J_i) admissible. In the preceding diagram only $\langle n, n \rangle$ is. Next, in analogy with the definition of an equilibrium point, let us define a (J_i) *point of justice* as a set of strategies, one for each player, such that adoption of these strategies leads to an admissible element as outcome.

The simplest justice-oriented rule of behavior is then:

(I) *If $J_1 = J_2$ and there is a unique point of justice, the strategy belonging to this point ought to be chosen.*

Unfortunately, (I) is not applicable to the prisoner's dilemma, for the requirement that there be a unique point of justice is not satisfied.[13]

A more complicated, but still relatively simple ethical rule of behavior may be introduced in terms of the notion of a *justice-saturated* strategy. A strategy for player i is justice-saturated (with respect to J_i) if whatever strategies are picked by the other players the resulting set of strategies is a (J_i) point of justice. The rule of behavior is then:

(II) *If for any player this set of justice-saturated strategies is non-empty, he ought to choose one.*

In the prisoner's dilemma each player has a unique justice-saturated strategy, namely, the strategy of not confessing, joint use of which leads to the reasonable outcome $\langle v, v \rangle$.

To be sure, when a person's set of justice-saturated strategies contains more than one element, (II) does not lead to a unique action, and some supplementary ethical rule of behavior may be needed. A similar problem arises for prudential game-theoretical rules of behavior and should

[13] It is perhaps useful to mention that in general a game of the type being considered here does not have a unique equilibrium point; the prisoner's dilemma is a happy exception.

surprise only those who believe that satisfactory categorical rules of action are easily come by.

If neither (I) nor (II) is applicable (and simple two-person decision situations exist for which this is the case), the theory of justice outlined here is of no use in determining what action to take, except insofar as the relation J_i is a constraint on the person's utility function.[14]

But this last problem of applicability is one of the least difficulties that face an adequate formal theory of justice. For example, even when (II) is applicable, an "ethical" man using it may be at a definite competitive disadvantage against a "prudential" man. In the prisoner's dilemma if prisoner I adopts his justice-saturated strategy and prisoner II his equilibrium-point strategy, then prisoner I will receive the maximum conviction. It is not easy, for me at least, to decide if this is an intuitive argument against the formal theory of justice or fair play set forth here, or if it is an intuitively reasonable instance of a just or fair man getting the worst of a situation. If the latter is the case, I think it may be claimed that a man who in all situations acts according to ethical rules of behavior may fare as well in the long run as the purely prudential man, provided knowledge of his standards of actions are known to his fellow man.

Another difficulty with the present theory is its structural weakness. It is a priori certain that no very elaborate theory of action can be built on the simple notion of a strict partial ordering. A major step in the development of rational theories of behavior has been the quantification of value (i.e., utility) and of subjective probability (i.e., reasonable degree of belief). Plausible assumptions which will lead to quantification of the theory of justice seem hard to find.

Making the theory of justice depend on the individual preference rankings is very much in the spirit of modern welfare economics, but may seem highly unsatisfactory to many philosophers. And I think it may be rightly objected that the intuitive success of the theory depends upon these individual preference rankings themselves satisfying certain criteria of justice. To admit this objection is not to accede to a charge of circularity, for moral principles of justice, logically independent of the theory developed here, can be consistently introduced as constraints on individual

[14] In general, finite games only have equilibrium points when mixed strategies (i.e., probability mixtures of pure strategies) are admitted. A discussion of Rules (I) and (II) with respect to mixed strategies would take us too far afield.

preference rankings of $C_1 \bigcup C_2$. I simply do not have at the present any such interesting formal principles to suggest.

However, it may be appropriate to mention an alternative way of treating the theory developed here. The one detailed application has been to a non-cooperative game. In a cooperative game, for instance, an arbitration situation, it might be reasonable for the two participants who are in conflict, but who are upholders of ethical rules of behavior, to appoint an arbitrator they both trust. The arbitrator is then asked to make what he considers the fairest preference ranking of $C_1 \bigcup C_2$ in terms of his knowledge of the participants' needs and wants. Rules (I) and (II), if applicable, might then determine the outcome of arbitration. The immediate objection to this seems to be that if the arbitrator is going to do the ranking, why not simply let him rank the outcomes, and then agree on the one he considers fairest as the negotiated outcome. There is a simple answer to such an objection. It may be easy to rank $C_1 \bigcup C_2$, but very difficult to rank $C_1 \times C_2$. For example, let

$$C_1 = \{\text{trip to Hawaii, trip to N.Y.}\},$$
$$C_2 = \{\text{trip to Florida, trip to Chicago}\},$$

and the arbitrator, knowing persons I and II, may find it easy to rank $C_1 \bigcup C_2$:

trip to Hawaii,
trip to Florida,
trip to N.Y.,
trip to Chicago,

but he finds it very difficult to compare elements of $C_1 \times C_2$ like <trip to Hawaii, trip to Chicago> and <trip to N.Y., trip to Florida>.

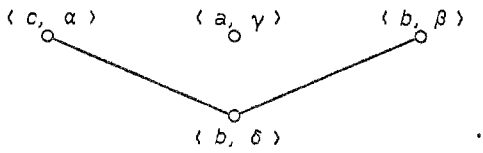
In conclusion an example may be constructed for which equilibrium-point analysis seems to lead to a more equitable and just solution of a non-cooperative game than the theory of justice outlined here. Let

$$C_1 = \{a, b, c\}$$
$$C_2 = \{\alpha, \beta, \gamma, \delta\},$$

let $R_1 = R_2 =$ the ranking: $a, \alpha, b, \beta, c, \delta, \gamma$, and let the game matrix be:

I \ II	1	2
1	$<a, \gamma>$	$<b, \beta>$
2	$<c, \alpha>$	$<b, \delta>$

Then $J_1 = J_2$, and we have as the Hasse diagram of the justice partial ordering of $C_1 \times C_2$:



It is easily checked that decision 1 is the unique justice-saturated strategy for each player, yielding the outcome $<a, \gamma>$, whereas the unique equilibrium point strategies yield $<b, \beta>$ as the outcome. In terms of the ordinal properties of the consequences at least, outcome $<b, \beta>$ seems fairer than $<a, \gamma>$. I conclude that the theory of justice developed here satisfactorily solves only a certain perhaps small proper subset of two-person, non-cooperative games.

The difficulties of formulating a theory of justice for even a very restricted set of situations suggests there may be something seriously wrong with this kind of effort, at least in terms of any principles we seem able to formulate at present. What seems needed as a prolegomena is the painstaking working out of some less sweeping, more concrete grading principles of the sort needed to take a position on particular issues of economic, political or social significance. Example 4 is a sketch of one sort in this direction.

REFERENCES

Blackwell, D. and Girshick, M. A., *Theory of Games and Statistical Decisions*, Wiley, New York, 1954.

Braithwaite, R. B., *Theory of Games as a Tool for the Moral Philosopher*, Cambridge University Press, Cambridge, 1955.

Hare, R. M., *The Language of Morals*, Oxford University Press, Oxford, 1952.

Kant, I., *Critique of Pure Reason*. Trans. by Max Muller, 2nd ed., rev. Macmillan, New York, 1949a.

Kant, I., *Critique of Practical Reason and other Writings in Moral Philosophy*. Trans. and ed. by L. W. Beck, University of Chicago Press, Chicago, 1949b.

Lange, O. and Taylor, F. M., *On the Economic Theory of Socialism*. Ed. by B. Lippincott, University of Minn. Press, Minneapolis, 1938.

Luce, R. D. and Raiffa, H., *Games and Decisions: Introduction and Critical Survey*, Wiley, New York, 1957.

Nash, J. F., 'The bargaining problem', *Econometrica* **18** (1950) 155–162.

Nash, J. F., 'Non-cooperative games', *Annals of Mathematics* **54** (1951) 286–295.

Savage, L. J., *Foundations of Statistics*, Wiley, New York, 1954.

Suppes, P., 'The role of subjective probability and utility in decision-making' in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* **5** (1956), pp. 61–73.