MORAL COMPLICATIONS AND MORAL STRUCTURES*

Robert Nozick

I. Introduction

In this essay I shall discuss some problems in representing one structure which may be exhibited by part of the moral views of some people. In particular, I shall be concerned with formulating a structure which would generate (or play a role in the generation of) a person's judgments about the moral impermissibility of specific actions. I do not claim that *everyone's* moral views exhibit the structure I shall discuss, nor do I claim that there is some one structure which everyone's moral views exhibit. Perhaps people's actual moral views differ in structure. Why I consider the particular structure I work towards worthy of discussion, given only the weak claim that *some* people's views may exhibit it (and perhaps only at a superficial level), will become clear as I proceed.

The discussion strikes what I hope is not too uneasy a balance between a descriptive and a normative interest in the structure of a person's moral views. I would view it as an objection that no correct moral view could have the structure I discuss. And I would hope that a person holding the type of view I discuss would find disconcerting the discovery that his view does not satisfy some of the specific structural conditions discussed later in the paper. Since people's views often do not satisfy conditions which they should satisfy, the imposition of normative structural conditions upon a person's moral views lessens the likelihood that people will be found whose views exactly fit the structure described. I would hope, however, that some people's views satisfy the structure closely enough so that it will be useful in accounting for the judgments about the moral impermissibility of actions that they actually make or would make. Thus the essay is intended to provide the beginnings of a model of a moral judge which is useful both for descriptive purposes and for various normative inquiries, in much the same way as is contemporary utility theory and decision theory.

Before proceeding, I wish to emphasize the exploratory nature of this

* Parts of this paper were delivered in an address given to the twelfth annual meeting of the Board of Editors of the Natural Law Forum, September 29, 1967.

essay, which contains little more than the first steps towards one moral structure. Since a major purpose is to open an area for investigation, I have felt free to mention in passing some questions without attempting to answer them here. Indeed, the last section includes a listing of further questions and problems which must be pursued if the kind of view suggested here is to be made to work. I would have preferred to present here a finished and complete work, and hope that the presentation of a temporal section of an ongoing work will have virtues of its own.

Though much has been written on particular problems within ethics, one finds, in the literature of moral philosophy, very little detailed discussion of the structure of a person's actual moral views. This may be due to its being assumed that a person's actual moral views exhibit one of two very simple structures (the first being a simple case of the second, which is worth mentioning separately). And this assumption may be made plausible by the desire, of many moral philosophers, to propose one or a small group of principles which would account for and unify a person's actual moral judgments (with a little fiddling here and there in the interests of consistency and theoretical simplicity).[1] If one or a small group of principles underlies all of one's moral judgments, then one will feel no need to discuss the structure of a person's moral views beyond stating the principles and saying something about the way in which the particular judgments are "derived" from the principles. For, it would seem, all of the interesting questions about a person's moral beliefs could be raised about the unifying principles.

The two structures which it is often assumed or explicitly stated that (part of) a person's moral views exhibit I shall call the maximization structure and the deductive structure.

A. *The Maximization Structure.* According to this view all of one's judgments about the moral impermissibility of actions are accounted for by a principle which requires the maximization of some quantity, subject perhaps to a quantitative restraint. The traditional utilitarian, or at any rate the traditional utilitarian of the textbooks, makes this claim. A proponent of such a view may discuss problems about knowing which act maximizes the quantity, or problems about what to do when one lacks adequate information to decide which act will maximize the quantity, but he will view all of the important questions about structure as obviously answered.

I do not wish here to discuss utilitarianism; I shall just assume that it is an inadequate moral view. Even though it is inadequate, some people

[1] An especially clear and self-conscious example of attempting to carry out this task is Henry Sidgwick's METHODS OF ETHICS. See the bibliography, pp. 49-50.

may hold it. If so, I am not trying to describe the basic structure of *their* moral view. However, if they are utilitarians because they believe this position accounts for their other moral views, it may be that the structure I discuss will capture their views at a less deep level. The inadequacy of utilitarianism, of course, does not show that every view with a maximization structure is inadequate. Perhaps some such view is adequate, and perhaps such a structure underlies and accounts for the different structure I shall discuss, in an interesting way. I say "in an interesting way" for it may always be possible to produce a gimmicky real-valued function such that its maximization mirrors one's moral views in a particular area. For example, suppose there is a complicated theory T, not having a maximization structure, which accounts for one's judgment of which actions are morally impermissible, and which are not morally impermissible. Define a function f taking actions as argument values, such that $f(A) = 0$ iff, according to T, A is morally impermissible; and $f(A) = 1$ iff, according to T, A is not morally impermissible. One might then say that (part of) one's moral views are represented by the requirement that one maximize (act so as to achieve maximal values of) f. An interesting maximization structure requires the maximization of some function which was not gimmicked up especially for the occasion (or similar occasions). That is, one that is interesting for our purpose; there may be other reasons why one might find interesting the representing of moral views via an artificially created maximization structure. Compare the situation with contemporary utility theory, which shows that if a person's preferences among probability mixtures of alternatives satisfy certain natural-looking conditions then his preferences can be represented as being in accordance with the maximization of the expected value of a real-valued function which is defined in terms of the conditions and his particular preferences (the utility function). One would not answer Aristotle's question of whether there is one thing for which (eventually) all actions are done by saying, on the basis of contemporary utility theory, "yes there is, namely, utility." To rule out, on theoretical grounds, such an answer, and to rule out similar maximization structures in ethics as interesting for our purposes, one would need a way of distinguishing the one from the many (gimmicked up to look like one). It is clear that the formulation of such a way faces problems similar or identical to the ones faced in attempting to rule out Goodman-like predicates,[2] and I shall not pursue this matter further here.

Many persons do not believe that there is an interesting maximization structure (and many believe that there isn't one), which accounts for their

[2] Nelson Goodman, Fact, Fiction, and Forecast ch. III (1955).

views about which actions are morally impermissible. But even if their moral views can (ultimately) be accounted for by an interesting maximization structure, it is still of great interest to see if one can state a structure (which would, by hypothesis, be reducible to an interesting maximization structure) which accounts for their moral views and which they can recognize as their own.

B. *The Deductive Structure.* According to this view, a particular judgment that an act is morally impermissible

1) Act A is morally impermissible

would be accounted for as following from statements of the following form:

2) Act A has features F_1 , ... , F_n

3) Any act with features F_1 , ... , F_n is morally impermissible.

I shall consider only the simplest case where the features F_i are empirical, factual, not *explicitly* moral features of actions. The moral premiss 3) is accounted for as following from the moral premiss

4) Any act with features T_1 , ... , T_m is morally impermissible

conjoined with the factual premiss

5) Any act with features F_1 , ... , F_n has features T_1 , ... , T_m.

It is not claimed that the person consciously draws such inferences, but it is claimed that, in some sense, such a structure accounts for his particular judgments, and that the person would be willing to say, or someone who understood his view would be willing to attribute to him the (perhaps implicit) belief that an act with features F_1 , ... , F_n is wrong because it has features T_1 , ... , T_m.[3]

Principle 4) is accounted for in the same way as is 3), by reference to another moral principle, and factual premiss. Presumably, one eventually ar-

[3] He might not say "because it has features T_1, ... , T_m", but rather "because it has features ——" and fill in the blank with some though not all of the *m* T-features. The task of accounting for the shortening of the answer is similar to that faced by an adherent of the deductive-nomological model as an adequate account of a certain kind of explanation (or as presenting necessary conditions for one important kind of explanation) in explaining why when one says "Event E happens because ——" one does not normally fill in the blank with *all* of the things that the model says forms part of the explanation, not even with all of the initial conditions. For a discussion of the kinds of principles underlying the shortening of explanation answers (which is relevant to the point with which this footnote begins), cf. S. Bromberger, *An Approach to Explanation,* in R. BUTLER (ed.), ANALYTICAL PHILOSOPHY, Second Series (1965), *Why Questions,* in R. G. COLODNY (ed.), MIND & COSMOS (1966), and H. L. A. HART and A. M. HONORÉ, CAUSATION IN THE LAW ch. I-II (1959). On the deductive-nomological model of explanation, see C. G. HEMPEL, ASPECTS OF SCIENTIFIC EXPLANATION (1965). Though I shall later in the text be discussing a model of the structure of moral views other than the deductive one, this alternative model also faces the task of accounting for shortened answers. The literature cited here is also relevant to that task for the alternative model, which I shall not attempt in this paper.

rives at one or more moral principles P of the form of 3) and 4), which underlie all of the person's particular judgments, and all other of his moral principles, and which are such that there are no other principles P′ of this form to which the person would assent or which the observer would attribute to the person as more basic ("Acts having the features P mentions are morally impermissible because they have the features mentioned in P′") from which (members of) P can be derived in the way that others are derived from P.

The deductive structure cannot easily explain how it is that a person's moral judgment of a particular act (often) changes as he learns additional facts about the act; e.g., he no longer judges the act morally impermissible. In such cases the facts previously known to the person were not sufficient for the truth of a judgment of moral impermissibility, and he did not have knowledge which instantiated the antecedent of an exceptionless moral principle. The expedient of maintaining the deductive structure hypothesis by claiming that in all such cases the facts known to the person do not instantiate the features mentioned in an exceptionless moral principle but instead provide inductive evidence for their realization, while explaining how judgments would change with new information, is obviously implausible.

A second difficulty with attributing the deductive structure as the superficial structure of some people's views is that these people are unwilling to state or assent to any or very many exceptionless moral principles.[4] Many such persons, at some time in their lives, explicitly accepted such exceptionless principles, gradually making them more and more complicated to fit more and more complicated cases. Then at some point they decided they couldn't state exceptionless principles which they were confident were correct *and* which would account for a wide range of their moral judgments. Perhaps reinforcing their lack of confidence in any exceptionless principles they were tempted to state, was the realization that if more than one such principle is stated, great care must be taken to ensure that in no possible case do they conflict.

Such a history, I imagine, would be very common among lawyers, to whom the difficulties of devising rules to adequately handle, in advance, all the bizarre, unexpected, arcane and complicated cases which actually arise, much less all possible cases, are a commonplace. The view that any laws a legislature will be able to devise will work contrary to their intention, or will work injustices, in some cases they hadn't foreseen arising or even con-

[4] That is, principles of the form of 3) and 4) where it can be decided, with little room for fiddling, on nonmoral grounds whether the features mentioned in the principle apply to an action. For other sorts of features the persons I am considering might be willing to assent to exceptionless principles, e.g., "any action which shows lack of love of one's neighbor is morally impermissible."

templated, needs no exposition by me in a journal read by lawyers. Awareness of the difficulties in formulating rules to handle all the cases which will arise often leads to talk of the role of judicial discretion in a legal system, and to the incorporation within legal codes of statutes dealing with the avoidance of evils, which do not attempt to specifically handle the possible cases.[5] Thus one would expect lawyers to be as modest about the purported exceptionless character of any moral rules they can devise, or about most that come down to them in their moral tradition, as they are about the product of centuries of intensive legal effort to devise and refine rules to govern conduct.[6]

In saying that the persons would not confidently (or be willing to) put forth exceptionless principles which they believe would yield the correct judgment in each particular case, I do not mean to imply that they would not want themselves and others to publicly state such principles as exceptionless, and perhaps enforce them as such. They might want this, and do it because though they believe that the principles do not yield correct judgments in each particular case, they think it better that the principles always be followed than that each person consider and decide whether the situation he faces is an

[5] Cf. The American Law Institute, MODEL PENAL CODE sec. 3.02 (1962). No rules are provided whereby it is to be determined whether "the harm or evil sought to be avoided by such conduct is greater than that sought to be prevented by the law defining the offense charged," and presumably this is to be decided by a jury.

[6] It is worth briefly noting one kind of theoretical solution to this particular reason for desiring judicial discretion. ("Theoretical" because it ignores limitations upon the time and energy of legislators and legislatures.) In those cases, which are not constitutional cases and which needn't be decided immediately, where, under the doctrine of judicial discretion, the judge should *use* his discretion (e.g., an obvious injustice is worked by the law as it "clearly" stands, the legislature did not foresee such a case and would not, one thinks, have intended the consequence required by the law in such a case, etc.), the judge finds the person guilty and throws the case back to the legislature. That is, he notifies the legislature that here is a case where he believes an obvious injustice is worked by the law as it "clearly" stands, and the legislature has a specified amount of time to pass new legislation in this area. If under the new law, the defendant's act is no longer an offense, he is then found innocent. If some act performed by someone prior to the promulgation of the new law is an offense (only according to the new law), he is not tried.

In the preceding paragraph, I consider only those cases where, under the doctrine of judicial discretion, the judge should use his discretion to avoid the defendant's being found guilty of an offense. In the cases where the law "clearly" holds the person innocent, many of the objections to ex post facto legislation would apply to a judge's using his discretion to find the person guilty. (Though perhaps one would sometimes wish it to be done. Cf. e.g., Riggs v. Palmer, 115 N.Y. 506, 22 N.E. 188 [1889].) Similarly, though less strongly, for cases where no particular finding is "clearly" required by the law as written. In cases where two parties are opposing litigants, a more complicated procedure would be needed.

Many detailed questions arise about the operation of a system such as the one sketched above which, in view of the theoretical character of the solution, are not worth pursuing here. The reason for mentioning such a "solution" is to point out that no analogous procedure is even theoretically available to us in the moral case. For, putting it roughly, our task of accounting for an individual's moral judgments is like the one of formulating the rules and principles by which, under the scheme, the legislature decides the cases thrown back to it, or the judge decides when an injustice would be worked by the nondiscretionary application of the law.

exception to the principle. They might think this latter alternative worse in the belief that far more often or far more importantly there will be cases where people, using their discretion, wrongly do not follow the principle than where, without using discretion, they wrongly follow it. But still the persons I am considering would not believe that the principles they *say* are exceptionless always yield the correct results.

The theme of the inadequacy of most exceptionless principles that one can state fits in nicely with recent writings on prima facie duties and rights, a subject whose importance was first emphasized by W. D. Ross.[7] My procedure will be to present a relatively simple structure in harmony with these writings, and then to consider some objections to, criticisms of, inclarities in and infelicities in this particular structure. A consideration of these leads to some suggestions about what a more adequate (though similar) structure would be like, which (or modifications of which), it is hoped, would be a component of a theory to account for the moral judgments of some of us about the moral impermissibility of actions. I shall be wholly occupied in this paper in working towards the more complicated structure, and shall not get to present it in detail here.

II. The Simple Structure

There are two open-ended lists of features of actions: W (for wrong-making) and R (for right-making). Members of W are denoted by w_1, w_2, ... ; subsets of W by W_1, W_2, ... ; members of R are denoted by r_1, ... ; subsets of R by R_1, R_2, ... If an action has some features on W, and no features on R, it is morally impermissible. If an action has some features on R, and no features on W, it is morally required (morally impermissible not to do it), or at least morally permissible.[8] One key fact, which I mention in passing though much should be written on it, is that neither W nor R is empty. Furthermore, these are exclusive sets.

Few morally interesting acts will have (of its features on either list) features on only one of the lists. Thus we need, in addition to a way of representing what features the person considers to be morally relevant (which the lists are supposed to do), a way of representing his judgments that some features on one of the lists outweigh or override some features on the other

[7] See his books, The Right and the Good (1930), Foundations of Ethics (1939).

[8] Whether the R-list should contain (and contain only) features such that an action having some of these features, and none on the W-list, is morally required, is a question that requires much discussion. What is important is that it contain features which may override features on the W-list. The host of complicated questions about how to state this, and its consequences, will be considered later.

list. We shall represent these outweighings or overridings by inequalities between sets of features. Thus $W_1 > R_1$ indicates that an action with no features on either list other than the members of $W_1 \cup R_1$ (all of which it has) is morally impermissible.[9] If $R_1 > W_1$ then an action with no features on the list other than the members of $W_1 \cup R_1$ (all of which it has) is morally permissible (and perhaps morally required). Our initial statements about the actions which, of the features on either list, have only features on one of the lists, can now be written as

$$W_1 > \phi$$
$$\phi < R_1$$

where W_1 and R_1 are any nonempty subsets of W and R, respectively.[10]

To provide a way of representing drawing-the-line problems, and to provide a manageable way of representing a multitude of similar judgments, we shall allow some of the features on the lists to contain variables (in addition to that variable contained in every feature, which ranges over actions, reference to which variable is always suppressed in this paper) ranging over positive integers, e.g., "leads to the death of n persons."[11] Suppose this feature is w_1 which is a member of W, and consider a specific member of R, r_1. One might have as an outweighing $w_1 > r_1$; $n > 2$. This means that $w_1 > r_1$ for $n > 2$, and does not entail that not $-(w_1 > r_1)$ for $0 < n \leqslant 2$. A feature may contain more than one variable (the variables being ordered within the feature), and an inequality may contain more than one feature containing variables. This suggests that in these cases we must consider, in our representation, ordered sets of features on each side of the inequality, with an ordered set of numerical constraints afterwards. We should note the possibility that

---

[9] This seems an appropriate place to informally explain some symbols which sometimes are used at different places in the text, with which some readers may be unfamiliar. $X \cup Y$ (the union of set X and set Y) is that set containing as members just those things which are members of X, or Y, or both. ϕ is the set without any members, and is referred to as the null set. "$X \subset Y$" is read "X is a (proper) subset of Y" and is true if and only if each member of set X is also a member of set Y, and some member of set Y is not a member of set X. "$X \subseteq Y$" is read "X is a subset of Y" and is true if and only if $X \subset Y$ or X and Y have exactly the same members. A selection set from a group of sets is a set containing exactly one member from each set in the group. "$x \in X$" is read "x is a member of the set X." "$(\exists x)$ (. . . . x . . . .)" is read "There is an x such that (. . . . x . . . .)," e.g., "$(\exists x)$ (green x)" is read "There is an x such that x is green" or "There is something which is green." "(x) (. . . . x . . . .)" is read "For all x, (. . . . x . . . .)," e.g., "(x) (green x)" is read "For all x, x is green." "____ $\rightarrow$ . . . ." is read "if ____ then . . . . ." As an example of how things fit together

$$(X)\ (X \text{ is a set} \rightarrow (\exists y)\ (y \in X) \text{ or } X = \phi).$$

is read: For all X, if X is a set then there is a y such that y is a member of X, or X is equal to the null set.

[10] If an action with members of R and no members of W may be morally permissible though not morally required, it is not clear how to interpret $\phi < R_1$.

[11] I shall ignore in this paper any need for variables ranging over all the rationals or reals.

the inequality may contain as a constraint an equation essentially involving more than one of the variables, e.g., $nm \geqslant 2t$. I shall not pursue here the details of how such complications are to be best represented formally.

III. Ordering of Features and Their Combinations

What structural conditions on inequalities over features[12] may legitimately be imposed? Do these yield an ordering and, if so, of what strength?

We have a notation for outweighing or overriding. Do we need to introduce an equality sign for the notion of exactly balancing? The metaphor of scale balances would suggest that we do, but our explanations thus far do not seem to allow room for such a notion. For what would it mean to say that W_1 exactly balances R_1? It seems that an action having, of its properties on either list, all and only the members of $W_1 \cup R_1$ will be morally permissible or morally impermissible. If it is morally permissible, then $R_1 > W_1$; if it is morally impermissible, then $W_1 > R_1$. Where is the room for an equality sign?[13] The further possibility one thinks of is that the system of principles is incomplete in that, for some particular act having some features on each list, it yields neither the judgment that the action is morally permissible, nor the judgment that it is morally impermissible. It would be misleading to represent such a situation by an equality sign.

A condition which is obviously desirable is

$$1) \quad R_i > W_j \rightarrow \text{not} - (W_j > R_i)$$

(We needn't, of course, separately state its equivalent contrapositive.) The connectedness condition

$$2) \quad (R_i)\ (W_j)\ (R_i > W_j \text{ or } W_j > R_i)$$

is surely too strong a structural requirement to reasonably impose upon a person's fragmentary moral views. We shall take it up again later when we come to consider complete moral systems.

In addition to 1), obvious candidates are various transitivity conditions. Let X, Y, Z, S, T range over subsets (proper or improper) of R, and of W.

[12] In this section "features" shall refer to features on the lists without variables, or specifications of features, with variables, on the lists (i.e., with constants substituted for each variable). If certain conditions are met, features with bounded variables (the bounds being represented by equations) may be included also. I shall not pursue here the obvious way to do this.

[13] The metaphor of balancing and outweighing leads one to want to consider a notion of exactly balancing, and to want to distinguish moral views which hold that in exact balancing situations the action is morally permissible, from those that hold that in exact balancing situations the action is morally impermissible. Though our explanation of the inequality sign does not leave room for this distinction, it may be desirable to pursue the possibility of other structures which allow it.

$X > Y$ is well formed only if either $X \subseteq R$ and $Y \subseteq W$ or $X \subseteq W$ and $Y \subseteq R$. Consider

3) $(\exists Z)\ (X > Z\ \&\ \text{not} - (Y > Z)) \rightarrow$

a) $(S)\ (Y > S \rightarrow X > S)$

b) $(S)\ (S > X \rightarrow S > Y)$

Intuitively, this seems reasonable. If X outweighs something that Y doesn't, then X has more weight than Y, and consequently outweighs everything that Y does, and Y is outweighed by everything that outweighs X.[14]

From 3), together with 1), follows

4) $X > Y\ \&\ Y > Z\ \&\ Z > T \rightarrow X > T$

Furthermore, one can define a notion of one set of features having more weight than another set from the same list:

$X >> Y =$ df. $(\exists Z)\ (X > Z\ \&\ \text{not} - (Y > Z))$.[15]

Given 3), $>>$ is irreflexive, asymmetrical, and transitive, and hence establishes a strict partial ordering of each list.[16]

But is 3) a legitimate condition to impose? Two reasons might suggest that it is not:

1. Different features may interact differently with features on the other list.

2. A feature may be strengthened or weakened, its weight increased or decreased, by something other than features of actions on the lists.

The second of these reasons we shall consider later, along with other reservations which stem from applying the condition to notions other than "overriding" and "outweighing" as these are explained later. Consider 4). Can't there be a case where $X > Y$, $Y > Z$, $Z > T$, yet $T > X$? So as not to be misled by our inequality sign, let us rephrase the question. Can't there be actions A_1, A_2, A_3, A_4, such that

a) A_1 has all the features in $X \cup Y$, and no other features on the list, and A_1 is morally permissible.

b) A_2 has all the features in $Y \cup Z$, and no other features on the lists, and A_2 is morally impermissible.

c) A_3 has all the features in $Z \cup T$, and no other features on the lists, and A_3 is morally permissible.

d) A_4 has all the features in $T \cup X$, and no other features on the lists, and A_4 is morally impermissible.

---

[14] Strictly speaking, since 2) is not imposed upon our fragmentary moral views, this intuitive argument for 3) should be put somewhat differently.

[15] In the absence of 3), one might wish to define $X >> Y$ as $(Z)\ (Y > Z \rightarrow X > Z)\ \&\ (Z)\ (Z > X \rightarrow Z > Y)\ \&\ (\exists Z)\ (X > Z\ \&\ Z > Y)$.

[16] For an explanation of this and other ordering terminology, see P. SUPPES, INTRODUCTION TO LOGIC 220-23.

The only cases (under 1. above, as opposed to 2.) which appear to fit this, which I can think of, are cases in which, e.g., T and X interact so as to produce *another* W-feature (not in T ∪ X) which also applies to A_4. But in such cases, it is false to say that the only feature on the lists which are had by A_4 are in T ∪ X.[17] Thus, I suggest that we tentatively accept condition 3), appropriately modifying it later when we come to discuss the difficulties under 2. Consider the further condition

$$5)\quad W_1 \subset W_2 \subseteq W \rightarrow$$
$$a)\quad (\exists R_1)\ (R_1 > W_1\ \&\ \text{not–}(R_1 > W_2))$$
$$b)\quad (\exists R_2)\ (W_2 > R_2\ \&\ \text{not–}(W_1 > R_2)).$$

Or, to remove the here irrelevant possibility that there may happen to be no set of R-features "between" W_1 and W_2, consider

$$6)\quad W_1 \subset W_2 \subseteq W \rightarrow (R_1)\ (W_1 > R_1 \rightarrow W_2 > R_1).$$

Intuitively, 6) says that adding more W-features to an action cannot make it any better. 5), which given the previous conditions and definitions is equivalent to $W_1 \subset W_2 \subseteq W \rightarrow W_2 >> W_1$, says that the more W-features, the worse (*ceteris paribus*). (Perhaps we should restrict this to acts which are not infinitely bad, if there be any which are. I ignore this complication here.) Once again, there is the possibility that some features which are in W_2 though not in W_1, so interact with features in W_1, in R_1, or with the facts of the situation (not represented by features on the lists) as to produce a new R-feature which was not previously present. But once again in this case, it is false that the *only* morally relevant features the act has are in $W_2 \cup R_1$ (for some particular R_1 which is a candidate for satisfying the antecedent of the consequent of 6), but not its consequent). So once again, until the complications under 2. above are considered and incorporated, we shall tentatively accept and use 5) and 6).

IV. Alternative Actions

According to the simple model presented in Part II, whether or not an action is morally impermissible depends only upon the features of the action, and upon the inequality among two sets of features which it has. Let W_X be the set of all and only those features of act X which are on the W-list; let R_X be the set of all and only those features of act X which are on the R-list. According to the simple model, act A is morally impermissible if and

[17] Very difficult issues arise here, related to some about explaining the notion of "intrinsically valuable" when there are causal or probability connections between features, which must be considered in a full exposition of the theory.

only if $W_A > R_A$. In arriving at a judgment about the impermissibility of an act A, one need not, according to the simple model, consider the alternative acts available to the person, nor need one consider longer courses of action of which A may be a part. The availability of alternative actions, and the embedding of an action in a longer course of action, produce complications which require modification of the simple model. In this section I shall consider only the first, leaving the second until Part VI.

Is it true that A is morally impermissible only if $W_A > R_A$? Is it true that its W-features' outweighing its R-features is a necessary condition for an act's being morally impermissible? If $R_A > W_A$, is this sufficient for A's being morally permissible?

Consider the following two situations, obtained from one often mentioned in the literature in discussions of lying.

1) You are present as Q flees down a road from P, who you know will unjustifiably physically harm or kill him. P comes running along and asks which way Q went. If you say nothing, he will continue along the road and catch Q. *The only way* to prevent this is to lie to P, telling him that Q went in a different direction. I assume that sufficient details can be filled in so that you will all agree that it is morally permissible to lie to P in this situation. The right-making feature of saving Q from great harm overrides the wrong-making feature of lying to P.

2) The same situation as in 1), except that now there is some other way to save Q from the harm which does not involve lying to P, or any other wrong-making feature, e.g., if you start to tell P that what he's doing is wrong, he'll stop and listen and be convinced and won't continue on after Q. I shall assume that in this situation it would be morally impermissible to lie to P.[18]

I claim that the action of lying to P in 2) has the same features as it does in 1). I shall say more about this claim below. Now I wish to explore its consequences. Since by hypothesis (in 1)) the R-features of this act outweigh the W-features, so do they in 2). Since the act in 2) is morally impermissible, this shows that the W-features of an act outweighing its R-features is not a necessary condition for that act's being morally impermissible.

What principles can we formulate to handle such issues? The simplest principle which suggests itself, to handle the case just described, is

I) If $W_A \neq \phi$ and $(\exists B)$ (B is an alternative action open to the person

[18] It does not matter whether you agree with my moral conclusions about these particular two situations. What is important is that you can find two situations with structures parallel to these which you would view as I view these: e.g., in 1) the only way to stop P is to shoot him; in 2) you can also stop him by lying to him.

& $W_B \subset W_A$ & $R_A \subseteq R_B$) then it is morally impermissible for the person to do A (even if $W_A < R_A$).[19]

Roughly put, I) comes to the claim that it is impermissible to do an action if one can achieve the same R-features at a cost of fewer W-features. I) says that you must not do an action if an alternative action enables you to achieve the same good at less cost. A parallel principle would require that one not do an action, with non-null W-features, if an alternative action enables one to achieve a greater good at the same cost. Thus

II) If $W_A \neq \phi$ & ($\exists$ B) (B is an alternative open to the person & $W_B \subseteq W_A$ & $R_A \subset R_B$) then it would be morally impermissible for the person to do A (even if $R_A > W_A$).

Note that II) does not require that a person maximize R-features, but requires only that he not pass up any *if* he is going to incur some W-cost. But still, II) seems to me to be too strong a condition to impose, and we shall *not* do so.

But we want something more than I). For though A may achieve more R-features than B, perhaps the extra gain of A isn't worth its extra cost. Two principles suggest themselves, the first being a special case of the second:

III) If ($\exists$ B) (B is available to the person and $W_B = \phi$ & $W_A \neq \phi$ & $R_B \subset R_A$ & $W_A > (R_A\text{-}R_B)$) then it is impermissible to do A (even if $R_A > W_A$).

IV) If ($\exists$ B) (B is available to the person & $R_B \subset R_A$ & $W_B \subset W_A$ & $(W_A\text{-}W_B) > (R_A\text{-}R_B)$) then it is impermissible to do A.

First a word about how to interpret III) and IV). If the moral world is very simple so that the total wrongness or rightness of a set of features is just the sum of the individual wrongness (rightness) and there is no interaction between the R and W features, then (X-Y) is just the set theoretic difference between X and Y; e.g., $(W_A\text{-}W_B)$ is just the set of features which belong to W_A and which do not belong to W_B. And the inequality sign, as before, appears between two sets of features. If however, as seems likely, things are not so simple, then (X-Y) must be interpreted as some numerical measure of the difference in value between X and Y (where, e.g., value is measured on an interval scale; i.e., a scale unique up to a positive linear transformation). And the inequality sign stands for the ordinary relation between numbers. I shall say something about how such numerical measures might be obtained in Part V. An important point for us to note here is that

[19] I am here assuming the truth of condition 5) of Part III. If this condition or some of the others accepted in Part III are inadequate then more complicated principles must be formulated here. But even if the conditions in the previous section are inadequate and must be replaced, we can adequately raise the issues we are concerned with here by using them.

some measure of by *how much* some features outweigh others, are better than, are worse than others, seems required to account for a person's judgments of moral impermissibility, even *apart from* considerations about "risk," as the utility theorists speak of it. This point is further reinforced by the condition VII, which follows below, which prevents assuming simple additive conditions without making them explicit, and prevents proceeding by speaking only of set-theoretical differences.

Thus far we have considered only those cases where $W_B \subseteq W_A$. But one wants principles to cover some cases where though A does not have each of the W-features that B has, the *other* ones which B has are less bad than A's. The simplest case of this sort is one in which B has at least all of the R-features that A does. This suggests

V) If $W_A \neq \phi$ & $(\exists B)$ (B is an alternative available to the person and $(W_A\text{-}W_B) > 0$ & $R_A \subseteq R_B$), then it is impermissible to do A.[20]

If V) is acceptable, one wonders why it is necessary to have $R_A \subseteq R_B$. Couldn't one have R_A and R_B be of equal weight, or R_B have greater weight than R_A, even though R_A is not a subset (proper or improper) of R_B? This would give us

VI) If $W_A \neq \phi$ & $(\exists B)$ (B is an alternative available to the person & $(W_A\text{-}W_B) > 0$ & $(R_B\text{-}R_A) \geqslant 0$), then it is impermissible for the person to do A.

But what of the cases where there is no B with less weighty W-features than A's which has at least as weighty R-features, though there is a B with less weighty W-features and R-features, and the extra gain in R-features from A rather than B is outweighed by the extra cost in W-features from A rather than B.

One might formulate the following principle to fit this:

VII) If $(\exists B)$ (B is an alternative available to the person & $W_A >> W_B$ & $(W_A\text{-}W_B) > (R_A\text{-}R_B)$) then it is impermissible to do A.

Consider the following intuitive argument for VII). Suppose there were an act C which just took up the W and R slack between B and A. Thus, B & C has just the same R and W features which A has. You've already decided to do B. Should you do C in addition? If the answer is no, it is impermissible to do B & C rather than B. Since W_C $(= W_A\text{-}W_B) > R_C$ $(= R_A\text{-}R_B)$, the answer is no. (I here suppose that the complication to be discussed in Part VI does not apply to act C.) Thus it is impermissible to do B & C. Since

[20] One does not need a scale of measurement stronger than that gotten in Part III, to state condition V. For $(W_A\text{-}W_B) > 0$ can be stated, using the notation of Part III, as $W_A >> W_B$. Similarly, one may state condition VI below without assuming such a stronger scale of measurement. However, the statement of VII, below, *does* require the assumption of a stronger scale of measurement than we yet have available.

B & C has exactly the same moral features as A, it is impermissible to do A.[21]

A stronger principle than VII is

VIII) If ($\exists$ B) (B is available to the person & $W_A \geqslant W_B$ & $(W_A\text{-}W_B) > (R_A\text{-}R_B)$), then it is impermissible for the person to do A.

VIII) differs from VII) only in the case where $W_A = W_B$. (And in this case we would have $R_B > R_A$.) Note that an intuitive argument for VIII), similar to the one for VII), can be offered only if the R-list is such that an action having some features on it and no features on the W-list is morally required. For an action C which, in this case, took up the R and W slack between A and B would have no W-features, and would have some R-features. If such an action C is morally required, we will have the intuitive argument for VIII). VIII) seems to me to be too strong a principle to impose. We shall thus have to be careful later not to construe the R-list so that all actions having some features on it, and no features on the W-list are morally required. If it is so construed, then principle VIII) must be admitted as legitimate.

I am more certain that one does not wish to require that a person maximize the difference between the R-weight and W-weight of an action. One does *not* wish to require

IX) If ($\exists$ B) (B is available to the person & $(R_B\text{-}W_B) > (R_A\text{-}W_A)$), then it is impermissible for the person to do A.

Note that IX) differs from VII) in that it does not have the clause that $W_A > W_B$ in its antecedent. Hence in a situation where $R_B = 52$, $W_B = 50$, $R_A = 2$, $W_A = 1$, IX) has the consequence that it is impermissible to do A, whereas VII) does not have this consequence. The intuitive justification for VII) was that A would be impermissible if its extra gain (over some other action) wasn't worth its extra cost (over this other action). This justification cannot be applied to IX) since IX) rules out actions which may have no extra cost over their alternatives.[22] Note further how an argument parallel to the one advanced for VII) gets blocked. One might consider an action C which takes up the moral slack between A and B. Thus $R_C = (R_B\text{-}R_A)$; $W_C = (W_B\text{-}W_A)$ where $W_B > W_A$.

[21] We might consider the stronger principle

VII') If ($\exists$ B) (B is an alternative available to the person & $W_A >> W_B$ & $(W_A\text{-}W_B) \geqslant (R_A\text{-}R_B)$), then it is impermissible to do A.

This principle, unlike VII), covers the case where $(W_A\text{-}W_B) = (R_A\text{-}R_B)$. Given a scale of measurement we might be able to interpret this, but we could not, if my remarks in Part III were correct, offer an intuitive argument for VII') similar to that offered for VII). For the action C which took up the difference between B and A would have to be such that $W_C = R_C$. And we have not seen any way to interpret this.

[22] Paralleling the economists' notion of opportunity cost, one may deny this. My claim, to be elaborated below, is that all morally relevant opportunity cost of an action is already implicitly built into its W-features.

The person has already decided to do A. Is it permissible for him to do C? The answer is yes, for by hypothesis (of IX) $R_C > W_C$. Thus it is permissible for the person to do A & C, and hence permissible for him to do B, if there is no other action which stands in the relation VII) describes to B. However, from this one cannot conclude that he is required to do C after doing A; hence one cannot conclude that he is required to do B rather than A. (One could conclude this only by imposing a much too strong interpretation of the lists and inequalities; viz., that if $R_X > W_X$ then it is morally impermissible not to do X.) VII) is the strongest principle of the sort we have been considering which seems to me appropriate. IX), it seems to me, must clearly be rejected, though the rejection of VIII) is more doubtful. But some complications about VII) (and the other conditions accepted) must be mentioned.

A. If there are an infinite number of actions available to the person, then for each one there may be some other action with the same R-features and less weighty W-features. For example, suppose there are an infinite number of alternatives $A_1, A_2, \ldots, A_i, \ldots$ and the measurement scales yield the results that (i) (j) $R_{A_i} = R_{A_j}$ and (i) $W_{A_i} = \frac{1}{i}$. There will be no action with least weighty W-features. Hence, according to VII), each of the acts $A_1, A_2, \ldots$ is impermissible. It is this sort of problem that leads writers in decision theory to speak of ϵ-optimizing. I shall assume that a similar line must be taken here, and shall ignore the details of how VII) is to be modified, and how the particular ϵ is to be chosen.

B. Suppose that in our original situation (of lying to save someone's life when some alternative also saves his life and involves less weighty W-features) the only other alternative which saves the person's life involves you in great personal expenditure of money, effort, energy, time, etc. In *some* situations where the alternative with less weighty W-features involves great personal cost and inconvenience (though not additional W-features due to this) one would not require that the person not perform the act with the weightier W-features.[23] Where A is the act with the W-feature, and B is an alternative which stands in the relation described by VII) to A, and B imposes personal (non-W) costs on the performer of B, how are we to decide whether or not A is permissible? Presumably this decision depends upon $(W_A\text{-}W_B)$, and upon how great the personal costs of B are (how much greater than they are with A.) But how exactly is a line to be drawn?

One might consider an act C which is like B except that the personal

[23] I am not claiming that it is always permissible to lie to someone to save oneself personal expense and inconvenience, or that it always is permissible when the R-features outweigh the W-features, but that it sometimes is in this latter case. No doubt, the pursuer's performing a wrongful act makes it especially clear in this case.

costs to the performer of B are imposed by C upon some neutral third party. C's imposing of these costs upon some third party presumably involves C in some W-features (in addition to those of B). The following principle — embodying the view that if you are required to throw the costs upon a neutral third party rather than do A, then you are required to take the costs upon yourself rather than do A — seems reasonable: If C stands in the relation to A described by VII), so that if C were available it would be impermissible to do A, then it is impermissible to do A (when B is available). That is, if the C corresponding to B is such that $W_C << W_A$ & $(R_A\text{-}R_C) < (W_A\text{-}W_C)$, then it is impermissible to do A when B is available and stands in the relation described by VII) to A, even though B involves personal costs to the performer.

Delicate questions arise over whether substituting "if and only if" for the initial "if" in this principle yields a legitimate principle; and if not, how the further cases are to be handled.[24] I shall not pursue these questions here, but shall hereafter suppose that we have answered them and have available to us a modified VII) which incorporates the above principle and whatever complications are needed to handle the further cases.

Let us denote by S the relation embodied in an adequately formulated VII) such that if some action available to the person stands in the relation S to A then it is impermissible for the person to do A. Thus we refuse to say that an action A is morally permissible if its R-features outweigh its W-features, but rather we, at this point, say that an action A is morally permissible if its R-features outweigh its W-features *and* there is no alternative action available to the person which stands in the relation S to A.

Let us consider one objection to this whole line of argument. It might be said that in the cases which prompted condition VII), act A has added features which are easily overlooked. For example, one might say that the act of lying to the aggressor to save a person's life when one can save the life by persuading the aggressor to stop, has the added feature (when compared to the act in the situation where the *only* way to stop the aggressor is by lying) of being an unnecessary lie. And, the objection would continue, if *this* W-feature is included, then $W_A > R_A$. So of course the action is impermissible. Hence, the objection concludes, there was no need to reject the simple structure. If $R_A > W_A$, this *is* sufficient for A's being permissible, and the whole line of argument in Part IV thus far has been mistaken.

[24] That is, are there cases where C does *not* stand in the relation described by VII) to A, yet it is impermissible to do A (if B is available) and if so, how are these to be marked off from the cases where B stands in the relation described by VII) to A, yet—because of the personal costs which B involves—it is not impermissible to do A?

We can rebut the particular "feature" which was suggested, by noting that "involves telling an unnecessary lie" is an explicitly moral notion, which explicitly evaluates comparatively features of acts. For notice that if he can stop the murder by shooting the potential murderer or by lying to him, we can say "he unnecessarily shot him; he could have lied to him," but not "he unnecessarily lied to him; he could have shot him." But, granting that the particular "feature" suggested won't do, isn't there some (other) way to capture "involves telling an unnecessary lie," *for a given situation,* which doesn't involve explicit reference to moral notions? And if so, won't this have to be ruled out as a feature of act A?

One might try, as such a feature of A

a) F= "is an alternative to another act B, available to the person, which has the same R-features as A and fewer W-features."

This obviously won't do since it explicitly mentions the notions of R-feature and W-feature. So one might try

b) F= "is an alternative to another act B, available to the person, which has features $w_1, \ldots, w_m$ and $r_1, \ldots, r_n$" where (though this is not said as part of F) $r_1, \ldots, r_n$ are exactly the R-features of A and $w_1, \ldots, w_m$ are a proper subset of the W-features of A.

Note first that even if this worked it would handle only the cases covered by Principle I) in this section. But it doesn't work, because B may have weighty features in addition to $w_1, \ldots, w_m$ which are on the W-list, and if so, F won't be a W-feature of A. (If one supposes F to be a W-feature of A in this case, *what R-feature(s) of A corresponds to B's having weighty W-features in addition to* $w_1, \ldots, w_m$, *and also overrides F?*) Trying to handle this possibility by adding into F "and B has no other features which are on the W-list" won't give us a non-explicitly-moral feature, and given the open-ended nature of the list one cannot build into F the conjunction which denies, for each other feature on the W-list, that B has it. Thus, it is not at all clear how to begin to state the candidate for the non-explicitly-moral feature (on the W-list) which A has when there is an alternative B to A which stands in the relation to A described by condition VII).

Thus, we continue to maintain that it is not sufficient for A's being permissible that $R_A > W_A$ (and thus we continue to reject the simple model), and we accept something like condition VII) as marking off the exception to the sufficiency. If there is available *one* other action of a certain sort, then A is impermissible, even though $R_A > W_A$.

Is it also the case that $W_A > R_A$ is insufficient for A's being impermissible, and do the same kind of objections to sufficiency hold as in the case of

$R_A > W_A$? Recall that we are tentatively accepting the view that

A) If $R_A > W_A$ then A is permissible unless
$(\exists B)\ (W_A >> W_B \ \&\ W_A - W_B > R_A - R_B)$. And if
$(\exists B)\ (W_A >> W_B \ \&\ W_A - W_B > R_A - R_B)$, then A is impermissible.

If things were symmetrical we would have correspondingly on the other side:

B) If $W_A > R_A$ then A is impermissible unless
$(\exists B)\ (W_B >> W_A \ \&\ W_B - W_A > R_B - R_A)$; and if
$(\exists B)\ (W_B >> W_A \ \&\ W_B - W_A > R_B - R_A)$, then A is permissible.

This says that if *one* action available is worse (in a certain way) than A, then A is permissible even though $W_A > R_A$. But this is absurd; an action is not permissible just because we could have done something else which was worse. Whereas it is not absurd (and is indeed true) to say that an action with non-null W-features was impermissible if we could have done something else which was, in a certain way, better. Thus, the situation is not symmetrical, and we have what I shall call the *First Asymmetry:* The existence of one alternative action of a certain sort makes $R_A > W_A$ insufficient for the permissibility of A; whereas it is *not* the case that the existence of *one* action of a certain sort makes $W_A > R_A$ insufficient for the impermissibility of A. *If* $W_A > R_A$ is insufficient for the impermissibility of A, this is not because of something about *one* of the alternatives to A.[25]

But is $W_A > R_A$ sufficient for A's impermissibility or not? It seems clear that if it is not, this is not because one of the alternatives to A is worse in a certain way, but can only be because *all* of the alternatives to A are worse in a certain way. And, indeed, it seems quite natural to say that when *all* of the alternatives to A are worse than A itself, then A is permissible even though $W_A > R_A$. Thus, one might say that

C) If $W_A > R_A$ then A is impermissible unless
(B) (If B is an alternative to A then
$W_B >> W_A \ \&\ W_B - W_A > R_B - R_A$); and if (B) (if B is an alternative to A, then $W_B >> W_A \ \&\ W_B - W_A > R_B - R_A$), then A is permissible even though $W_A > R_A$.

Against this one might plausibly argue that in each of the cases one thinks of which seem to fit $W_A > R_A$ yet *all* of the alternatives to A are worse than A itself, there are features which A itself has which capture the respects in

[25] This First Asymmetry thesis has the consequence that there is no duality which involves only the substitution of "permissible" for "impermissible," and "<" for ">," as there is in the simple structure with which we began in Part II. Talk of asymmetries brings to mind the famous asymmetry between good and bad; viz., that all good things must come to an end.

which the alternatives are all worse, e.g., "prevents the death of n-persons," "saves the lives of these other persons." And since in these cases A has the additional R-feature of avoiding or preventing certain evils, it is in fact false that its W-features outweigh its R-features, so that we are not faced with a case where $W_A > R_A$, and so are not presented with a counterexample to the claim that $W_A > R_A$ is sufficient for the impermissibility of A.[26]

If this counterargument is defensible, then we are faced with a *Second Asymmetry*. From our current vantage point, the First Asymmetry says that unlike $R_A > W_A$, where *one* action of a certain sort makes it insufficient for A's being permissible, one action cannot make $W_A > R_A$ insufficient for the impermissibility of A, and *if* anything about the alternatives to A does make $W_A > R_A$ insufficient for A's being impermissible, it is something about *all* of the alternatives to A. The Second Asymmetry claim says that *nothing* about the alternatives does make $W_A > R_A$ insufficient for A's being impermissible.[27]

Since something about the alternatives can make A impermissible even though $R_A > W_A$, the Second Asymmetry claim adds considerably to the significant divergence noticed by the First Asymmetry. (Of course, the simple structure of Part II was perfectly symmetrical.)

Is the Second Asymmetry claim true, or can A be permissible, even though $W_A > R_A$, because all of the alternatives to A are worse than A itself?

Let us consider a specific example.[28] A person is in the cab of a locomotive and approaching a three-way continuation of the route. If the train continues to go straight ahead, which it will do if nothing is done to its controls, it will run down and kill 20 people. If it is made to go on either the rightmost track or the leftmost track, it will in each case run down and kill 40 people (for each track has 40 people tied on it). All this is known to the person in the cab.

If the person allows the train to go straight ahead when there is *no one* on the side tracks, the act is wrong for it allows 20 people to be killed, and has no redeeming virtue. In the first described situation we think it permissible to allow the train to continue straight on because the W-feature of

[26] Note that it does not follow from this that in every possible situation there is at least one action available to a person which is morally permissible. It *may* be that a person can (wrongfully) intentionally put himself into a situation such that none of the alternatives in the situation he got himself into are morally permissible for him to do.

[27] Notice that we are here speaking of whether something *about the alternatives* to A can prevent A from being impermissible even though $W_A > R_A$. In Part VI we shall consider whether A can be morally permissible, even though $W_A > R_A$, because of something about larger courses of action of which A is a part.

[28] I owe this example, which led to a rewriting of an earlier version of this part of this section, to Professor Judith Thomson of The Massachusetts Institute of Technology.

allowing 20 people to be killed is outweighed by the R-feature of avoiding the killing of 40 people.[29]

There is one notion of "avoids" which involves reference to changing things from how they would be in the normal and expected course of events, and it is this notion which lends plausibility and punch to the Second Asymmetry claim. The example that we have been considering leads us to think that there is another notion of "avoids," where, roughly, the fact that all actions but one lead to a certain consequence is sufficient to yield the conclusion that that one avoids this consequence. But this additional "feature" of avoiding an undesirable consequence, in this sense of "avoiding," saves the Second Asymmetry claim only via the course of *trivializing* it. For of course all of A's alternatives being worse than A won't make A permissible, where $W_A > R_A$, if the fact that *all* these alternatives are worse creates an additional and very weighty R-feature of A (viz., avoiding .....) so that when this is included, $R_A > W_A$.

At this point we have a choice. We can say that the Second Asymmetry claim is true, though trivial. Or we can refuse to admit this sense of "avoid" as specifying a *feature* of an act for our purpose.[30] And in this case, while accepting the First Asymmetry, we might try to specify a *duality* thesis, which would involve the substitution of "$<$" and "$>$," "impermissible" and "permissible," and of universal and existential quantifiers, and would have the consequence that A and C above are duals. This latter alternative is certainly well worth pursuing, but in the absence of special reasons for doing so, and since it would make Part VI even more complicated, we tentatively choose the first alternative.

V. Measurement of Moral Weight

Those arguments in the previous section which depend upon conditions or principles which assume a method of *measuring* the difference in weights between sets of features on one of the lists may seem to the reader to be, though interesting and intuitively correct, merely useless speculation in the absence of the description of some method for obtaining such measurement. Such a reaction seems to me to be unduly harsh, but I shall not dwell on this since I wish now to sketch a method for obtaining the appropriate measurements.

[29] This is not the same as saying that if the train were proceeding along a side track it would be permissible to switch it to the center track because avoiding the death of 40 persons outweighs killing 20 persons.

[30] We might say that a feature of an act, for our purposes, is about *that* act, and not about, even implicitly, others. Of course, the problems in explaining *this* are immense.

The obvious suggestion is to consider probability mixtures of W-features, and probability mixtures of R-features, and to attempt to parallel the Von Neumann-Morgenstern or similar axioms for utility measurement[31] so that we get the measurement we need. This suggestion might be reinforced by noting that an adequate theory of moral judgment will have to account for judgments of actions under situations of "moral risk" (where associated with an action is a probability distribution over sets of features on the lists, rather than just one set of features on the lists), not to mention "uncertainty." So, it might be asked, why not introduce that apparatus at this point? It seems to me that this approach is inappropriate for our problem here. For our problem is not one of accounting for judgments in situations of "moral risk," and there seems to be no intuitive reason for introducing apparatus based upon probability considerations to handle the problem we are now faced with. If we were to utilize the VN-M-type of measurement, arguments would have to be offered to show why the numerical values thereby obtained should function in the principles we have already listed. The more desirable course seems to be to find a method of measurement utilizing only considerations intrinsic to the sort of situation in which our problem arose, or utilizing only that apparatus which is sufficient to generate the problem and to show the need for a method of measurement. Such an alternative course has an additional theoretical advantage. For if one can establish the existence of numerical scales, assigning numbers to sets of features, to be used in the principles for the situations discussed in Part IV, and if one can use a VN-M-type procedure to establish numerical scales, assigning numbers to sets of features, based upon and to account for judgments of "moral risk" actions, then one can raise the question: What is the relation between these scales? I shall not attempt here to pursue or even specify the issues of interest which might arise.

Is there some way, other than by a VN-M-type procedure, to establish the existence of numerical scales without utilizing information of a sort not already provided by the kind of apparatus we are discussing? If there is not, then there would be reason for believing that the sort of apparatus we are discussing is (if not supplemented) seriously inadequate. I should here like to sketch (and I shall here do no more than sketch) a way of obtaining the

[31] Cf. J. Von Neumann and Oscar Morgenstern, Theory of Games and Economic Behavior, 2nd ed., appendix (1953); R. D. Luce and H. Raiffa, Games and Decisions ch. II; G. Debreu, *Cardinal Utility for Even-chance Mixtures of Pairs of Sure Prospects,* 26 Review of Economic Studies 174-77 (1959). For weakening of the strong conditions, relevant to attempts to parallel them in the moral case, see M. Hausner, *Multidimensional Utilities,* in R. M. Thrall, C. H. Coombs, and R. L. Davis (eds.), Decision Processes, and R. J. Aumann, *Utility Theory without the Completeness Axiom,* 29 Econometrica 445-62 (1962).

numerical values which function in the principles of the previous section.

The procedure I suggest is very simple. Let me first state it in a way which will look circular. I hope that I have offered sufficiently forceful intuitive arguments for Principle VII in the previous section so that you will agree that we can assume that if there *were* numerical values, Principle VII would be operating. We can use this assumption to obtain, for a specific person, specific inequalities between differences. Without entering into the intricacies of the procedures to be used to discover this, I shall assume that we can determine that for the person some R-features outweigh or override some W-features, some W-features are worse than others, and some R-features are better than others. Suppose, for example, that we discover that (for the person)

a) Some R-features (call them R_A) outweigh some W-features (call them W_A).[32]

b) Some R-features (R_B) outweigh some W-features (W_B).

c) The W-features W_A are worse than W_B

d) The R-features R_A are better than R_B

Thus we have

a′) $R_A > W_A$

b′) $R_B > W_B$

c′) $W_A >> W_B$

d′) $R_A >> R_B$

If now we can find an action A whose only morally relevant features are exactly those in R_A and W_A, and an action B whose only morally relevant features are exactly those in R_B and W_B, *and* the person judges that it is morally impermissible to do A (if B is an alternative), then we can conclude that (for this person)

$$(W_A - W_B) > (R_A - R_B)$$

If we ask him questions about other combinations of features and situations such as the previous one, we will get additional inequalities between differences. The important fact is that if we can get enough such inequalities then we will have sufficient information to establish a numerical scale of a given strength.

Let me describe things somewhat differently. If the person is following Principle VII, which utilizes numerical values (if his judgments can be accounted for by this principle), then certain conditions will have to be satisfied.

[32] Throughout a)-d′) the labels of the form F_X and W_X are merely labels for sets of features and do not say anything more than this. I have used these labels, in a notation which is by now familiar because I shall soon discuss the sets of features for situations where the here arbitrary labels mean what they have meant previously.

For example, if these numerical values are to be measured on a scale which preserves relations among real numbers unique up to a positive linear transformation (an interval scale), then since it is a truth of arithmetic that

> If $X > Y > Z >$ zero, and $W > T > U >$ zero, and $X - Y > W - T$, and $Y - Z > T - U$ then $X - Z > W - U$,

if the numbers assigned to sets of features satisfy the antecedent, they must satisfy the consequent. So in particular, it will be true that

> If $W_A > W_B > W_C > 0$, and $R_A > R_B > R_C > 0$, and $(W_A - W_B) > (R_A - R_B)$, and $(W_B - W_C) > (R_B - R_C)$ then $(W_A - W_C) > (R_A - R_C)$.

Now this will be true only if the following is true. (Thus the following statement is a necessary condition for the weights to be represented on a certain kind of scale.) If (for a person) $W_A >> W_B >> W_C$ and $R_A >> R_B >> R_C$ and $R_A > W_A$ and $R_B > W_B$ and $R_C > W_C$ and the person judges that it is impermissible to do A in a situation in which B is available, and the person judges that it is impermissible to do B in a situation in which C is available, then the person judges that it is impermissible to do A in a situation in which C is available.

Note that this does not utilize any apparatus or notions beyond those we already had. Furthermore, it seems intuitively reasonable; that is, it is a condition one would wish to impose, and could well have imposed apart from all considerations about measurement.[33] We thus have a necessary condition for the existence of a measure of the moral weight of a set of features on an interval scale (I here implicitly assume, as throughout, that if there is a measure, Principle VII operates) which utilizes no very strong apparatus in its statement, and which furthermore seems intuitively reasonable and justifiable. One might hope to gather a large number of such necessary conditions, and prove that they are sufficient to establish the existence of an interval scale measuring moral weight. That is, for some locution already available to us from our previous apparatus, one introduces an n-place relation. The intuitively justifiable statements using this locution are written down as conditions on the n-place relation. Thus, corresponding to each intuitively justified statement (or rather, to a selection of these) is a condition on the n-place relation. If one has chosen wisely or luckily, one may then

[33] Perhaps the person could reasonably judge that, under the described circumstances, a particular A is impermissible if B is available for reasons quite different from the sort we are here considering. One should incorporate conditions excluding such reasons into the antecedent of the principle. (I assume there are a small number of kinds. One does not want to require in the antecedent of the principles the proper kind of reason unless there is a way of specifying it which does not require reference to an apparatus beyond the one we had before numbers were introduced.) In this way one can hope to eliminate any counterexamples to the principle which there may be, which are irrelevant to our concern here.

be able to prove, using these conditions on the n-place relation as axioms:

A *Representation Theorem:* showing that there exists a real-valued function assigning numbers which is such that, for specified numerical relations, these relations hold among the numbers if and only if some corresponding relation about the subject matter holds among the objects the numbers are assigned to; and

A *Uniqueness Theorem:* showing that any two real-valued functions shown to exist by the Representation Theorem stand in a certain relationship to each other. The more limited this relationship, the stronger a scale of measurement one has obtained.[34]

To remove one simplification in this sketch: it will not be the case that each of the conditions on the n-place relation which are jointly sufficient (it needn't be that each of them is necessary) to establish the existence of a measuring function will correspond to an intuitively justifiable normative condition or one which specifies the notions involved. For, if similar previous results are any guide, one will require in addition various structural conditions.[35] One hopes to find structural conditions, which when combined with the others will suffice for the task and which look as though (without too drastic an idealization) they are satisfied.

The detailed technical task of specifying the axioms which seem reasonable in the moral context, and which yield the result, I must leave for another occasion. Here I wish to point out that the prospects are very promising. One finds, in the literature, several axiom systems which either explicitly are about (or can be interpreted to be about) a 4-place inequality relation between differences. Which set of axioms one uses determines the strength of the scale one obtains (ratio, interval, higher ordered metric, ordered metric, etc.). Exactly which system of axioms should be adopted for our purposes here is a tricky question. Here I wish to confidently conjecture that some not very radical modifications of an already existing measurement system will capture intuitive moral conditions, (plus some structural ones) and will suffice to yield some reasonably strong measurement of the moral weight of a set of morally relevant features.[36]

[34] This is all put roughly. For detailed discussions of Representation and Uniqueness Theorems see P. Suppes and J. L. Zinnes, *Basic Measurement Theory,* in R. D. Luce, R. Bush, and E. Galanter (eds.), 1 Handbook of Mathematical Psychology (1963).

[35] On structural conditions, see P. Suppes, *Some Open Problems in the Foundations of Subjective Probability,* in R. E. Machol (ed.), Information and Decision Processes 162 (1960), and D. Scott and P. Suppes, *Foundational Aspects of Theories of Measurement,* 23 Journal of Symbolic Logic 113-28 (1958).

[36] For further details on the sort of existing system I am thinking of, see S. Siegel, *A Method for Obtaining an Ordered Metric Scale,* 21 Psychometrika 207-16 (1956); P. Suppes and M. Winet, *An Axiomatization of Utility Based on the Notion of Utility Differences,* 1 Journal of Management Science 259-70 (1955); and especially R. D. Luce

Before leaving the subject of measurement I wish to remove in advance two reasons that one might have for believing that the project sketched cannot be carried out successfully.

1. The modified Principle VII) of Part IV was more complicated than I have made it out to be in this section. Won't these complications interfere with the procedure of measurement? They will not if one is careful to construct the scale via the person's judgments for situations where

a) there is no extra personal cost to the agent in doing B rather than A
b) there is not an infinite set of available actions such that no action has least wrong W-features.

2. Does the project of measurement described above depend upon the claim (which the reader may consider not to have been established by my arguments in Part IV) that "act A has as an alternative action available to the person to whom A is an alternative, an act B which ....." is not useable as a feature of act A? It does not. For suppose that my claim in Part IV is mistaken. Then, if there is such an alternative act B available to the person, A has an additional W-feature so that $W_A > R_A$; e.g., the W-features of lying to the pursuer to save the other man's life would outweigh the R-features of this act (if there was a suitable nonlying alternative available). Still, given a reasonable independence assumption, the method sketched will yield inequalities of the form $W_A' - W_B > R_A - R_B$, where W_A' is the set containing all W-features of A except the ones which refer to the availability of a suitable act B as an alternative to A. And *these* inequalities between differences can be used, as the others before, to establish a numerical measure over subsets of features not containing features which refer to the availability of an alternative action.[37] Thus, the supposition that the claim in Part IV is mistaken does not block the proposed method of measurement.

VI. Larger Courses of Action

We have argued that for an act A such that $W_A \neq \phi$, $R_A > W_A$ is not sufficient for A's being morally permissible. For there may be available an alternative act B which stands in the relation to A described by condition VII) of Part IV, and in this case, A is not permissible. We have further argued that the claim that $W_A > R_A$ is a sufficient condition for A's being

and J. D. Tukey, *Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement*, 1 Journal of Mathematical Psychology 1-27 (1964).

[37] If we let W_A'' be the set theoretic difference $(W_A - W_A')$ then, given an additivity assumption which is not unreasonable for these cases, one will get (for the case where $W_A > R_A$) the result that $W_A'' > R_A - W_A'$. And by using this, we may get a measure of the weight of the "features" which refer to the suitable alternative acts.

impermissible cannot be similarly overthrown by considering situations in which all of the alternatives to A are worse than A. Should we conclude from this that $W_A > R_A$ is a sufficient condition for A's being impermissible, or are there some other considerations which yield the result that A may be morally permissible even though $W_A > R_A$?

The sort of situation which suggests itself is one in which though $W_A > R_A$, A is (a necessary) part of some larger course of action B, and $R_B > W_B$. For some such situation, may it not be that A is morally permissible? It might be suggested that this problem can be avoided, because in such cases "is a part of a larger course of action which is such that ....." will be an R-feature of the A, and hence it won't be true that $W_A > R_A$. Even if this is so (I shall take up the question later), one wants to know the appropriate way to fill in the blank. I shall approach this question by considering what conditions are appropriate supposing that such things are *not* features of acts. After doing this we shall then consider whether these conditions can be used to obtain an appropriate way of filling in the blank.

Let me give two examples of cases for which it might be said that though $W_A > R_A$, A is permissible because it is a necessary part of some larger course of action B, where $R_B > W_B$.

1) A person P is unjustly being pent up by another person Q. You steal from some innocent third party R, one key to the door, making it possible for you to release P. I assume that this act is permissible if and only if it is part of the larger course of action of obtaining the release of P. If, for example, you go on to throw away the key, sell it, put it in your scrapbook, then your stealing of the key (and not attempting to release the person) was impermissible. I thus assume that $W_{\text{stealing the key}} > R_{\text{stealing the key}}$, and thus that either "making it possible for you to release P" either isn't an R-feature of the act or, if it is, doesn't when combined with the other R-features of the act, override its W-features. But even though $W_{\text{stealing the key}} > R_{\text{stealing the key}}$, it may be permissible to steal the key.

2) A group of officials torture some person they know to be a terrorist, in order to discover the plans (which they know are about to be executed) of a terrorist group which they can then thwart, thereby saving many innocent lives. Assume that the officials are good, the terrorists are bad, and that saving these lives outweighs torturing the person. When the torturing is part of the larger course of action of saving the lives, it is permissible. If, however, the officials torture the person, obtain the information, and then do nothing with it, or just file it away,

then the torturing was impermissible. Thus $W_{torturing} > R_{torturing}$, yet the torturing may be permissible if it is a necessary part of a larger course of action B where $R_B > W_B$.

Thus it seems that $W_A > R_A$ is not a sufficient condition for the impermissibility of A, for A may be a necessary part of a larger permissible course of action. Can it be, on the other side, that parallel considerations about larger courses of action also prevent $R_A > W_A$ from being a sufficient condition for A's being permissible?[38] One thing which might suggest itself is that though $R_A > W_A$, A is impermissible because for every larger course of action A^1 of which it is a part, $W_{A^1} > R_{A^1}$. This seems to me to be a possibility not worth taking seriously. For it is difficult to believe that such a situation would not be reflected in the W-features of A itself.[39] But perhaps though A is part of some larger acts C which have $R_C > W_C$, each one of these has as an alternative an act (not containing A) which stands in the same relation to C as B stands to A in condition VII) of Part IV.

We want to say roughly the following:

1) $R_A > W_A$, but A is not permissible because ($\exists$ B) (B stands to A in the relation described by Principle VII) and *B is permissible*).
2) $R_A > W_A$, but A is not permissible because all courses of action of which A is a part are impermissible by 1).
3) $W_A > R_A$, but A is permissible because it is part of a larger B (where $R_B > W_B$) and B is permissible.

We cannot stop here in explaining the exceptions to

a) $R_A > W_A$ being sufficient for A's being permissible
b) $W_A > R_A$ being sufficient for A's being impermissible

for each of 1) – 3) has the word "permissible" after the "because," and it is the application of this word we are trying to account for. And we cannot eliminate "permissible" or "impermissible" after the "because" by rewriting 1) –3) and substituting "$R_X > W_X$" for each occurrence of "X is permissible" after the "because" and substituting "$W_X > R_X$" for each occurrence of "X is impermissible" after the "because." For our problem is just that these are not sufficient conditions for permissibility and impermissibility respectively, and we cannot state principles governing the exceptions to them as sufficient conditions which assume that there is no such problem. For example, the

[38] I have already argued that because of considerations about *alternative* courses of action taking up the same time interval as A does, $R_A > W_A$ is not a sufficient condition for A's being permissible.

[39] Suppose, for simplicity, a finite number of possible larger courses of action of which A is a part, A^1, A^2, ... , A^n. Then if "is part of an act which _____" *is* a feature of an act, then A would have as a W-feature "must be part of an act with features W_A1 or with features W_A2 or . . . . . or with features W_An."

act B referred to in 1) must not be one shown to be impermissible by 2); 2) explicitly refers to 1); the act B referred to in 3) must not be one shown to be impermissible by 1) or 2), etc. One gets complications piled upon complications — though fewer than if we had taken the course of rejecting the Second Asymmetry claim at the end of Part IV.

How might these complications be handled? Let us first define some notions.

1) B *undercuts*$_0$ A = df. A and B occupy the same time interval and $R_A > W_A$ & $R_B > W_B$ & $W_B << W_A$ & $(W_A - W_B) > (R_A - R_B)$.
2) B *strongly undercuts* $_t$ A = df. B is an act over the time interval t, and B does not contain A as a part, and no part[40] of B comes before A, and no part of A comes before B, and (X) (X is an act over t containing A as a part → B undercuts$_0$ X).
3) A *is strongly undercut* $_t$ = df. ($\exists$ B) (B strongly undercuts $_t$ A)[41]
4) C *begins a strong undercutting* $_t$ *of A* = df. ($\exists$ B) (B contains C & C begins B & B strongly undercuts $_t$ A).
5) A *is strongly undercut* $\geqslant_t$ = df. (t′) (t′ is an interval beginning when A does and extending at least up to t → ($\exists$ B) (B strongly undercuts $_{t'}$ A))
6) C *begins a strong undercutting* $\geqslant_t$ *of A* = df. (t′) (t′ is an interval beginning when A does and extending at least up to t → ($\exists$ B) (B contains C & C begins B & B strongly undercuts $_{t'}$ A))

I now want to define a special kind of strong undercutting $\geqslant$ t of A. Roughly, it is one, begun by some C, which gives at least *one* course of action continuing through the various time periods $\geqslant$ t, such that each segment (continuing up to t) of this *one* course of action strongly undercuts A. I do not see any way to define this notion using only the apparatus of first-order quantification theory.

Suppose C begins a strong undercutting $\geqslant_t$ of A. Then for each time period t_i beginning when A does and extending at least up to t, there is at least one action B which contains C as its beginning and which strongly undercuts $_{t_i}$ A. There may be more than one such action. Let S_i be the set of all such actions; i.e., where t_i is a time period beginning when A does and extending at least up to t, S_i = the set of all actions which contain C as their beginning and which strongly undercut $_{t_i}$ A. We now want to define

[40] Strictly, no part of measure non-zero. I shall omit this in what follows.

[41] One can define an apparently weaker notion: A is *weakly undercut*$_t$ = df. (X) (X is an act over t containing A → ($\exists$ Y) (Y does not contain A & Y undercuts$_0$ X)). Similarly one can get definitions paralleling the ones which follow by using "weakly undercuts" rather than "strongly undercuts" (and other obvious changes). Interesting questions arise about the relation of the "weak" notions to the "strong" ones.

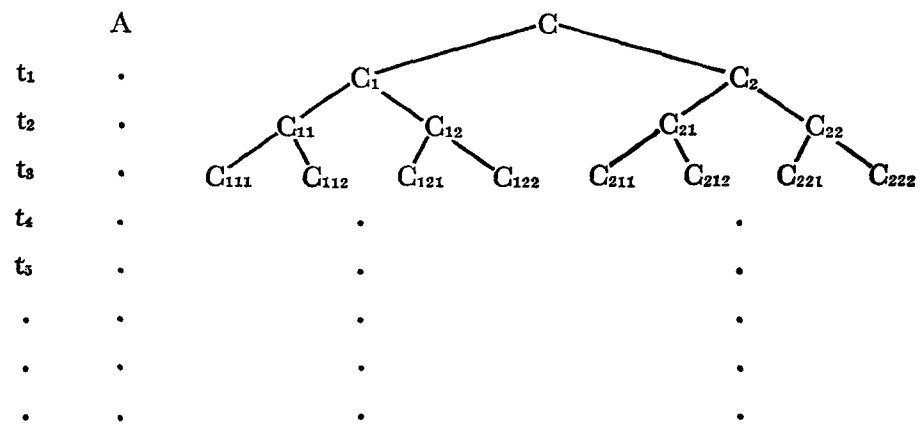
the one course of action spoken of above, which will be represented by a selection set from the family of the S_i (a set containing one member from each of the S_i). There will be one such course of action if and only if there is a selection set which represents one continuing course of action. Let us denote by B_i, the member of this selection set coming from the set S_i, and denote the selection set itself by S. S will represent the one course of action we want if and only if [t_i begins the interval t_j if and only if B_i begins the course of action B_j]. I assume that each time interval begins when action A begins. We define:

7) C *begins a sequential strong undercutting* $\geqslant_t$ *of A* if and only if C begins a strong undercutting $\geqslant_t$ of A and there exists a selection set S from the family of the S_i such that (B_i) (B_j) $[B_i \in S \ \& \ B_j \in S \rightarrow (t_i \text{ begins the interval } t_j \equiv B_i \text{ begins the course of action } B_j)]$.

Given these definitions, we may now state:

Principle I: If $R_A > W_A$ & $(\exists C)$ $(\exists t)$ (C is available to the person and C begins a sequential strong undercutting $\geqslant_t$ of A) then A is impermissible.

Can the antecedent of Principle I be weakened so as to yield another valid principle? If we eliminate the word "sequential" in Principle I, we get a principle which would hold A impermissible in the following sort of situation, where downward paths represent courses of action.



Suppose C undercuts$_0$ A. C & C_1 strongly undercuts$_1$ A, but C & C_2 does not; C & C_2 & C_{21} strongly undercuts$_2$ A, but no action over t_2 of which C & C_1 is a part strongly undercuts$_2$ A; C & C_1 & C_{12} & C_{121} strongly undercuts$_3$ A, but no action over t_3 of which C & C_2 & C_{21} is a part strongly undercuts$_3$ A, and so forth. Thus in this case, C begins a strong undercutting $\geqslant 0$ of A, but there is no t such that C begins a *sequential* strong undercutting $\geqslant_t$ of A. In this case, there is no one course of action one can recommend to the person as an alterna-

tive to beginning his course of action with A. That A is impermissible in this situation seems doubtful.

If in the previous example, node C_2 and everything that follows it is eliminated, then we get something between strongly undercutting $\geqslant_t$ and strongly undercutting$_t$; namely, cyclical undercutting. For in this case, for each odd i, C & C_i, begins a strong undercutting$_t$ of A, whereas this is not the case for even i. I shall not bother to define precisely the notion of a cyclical undercutting, or of a sequential cyclical undercutting. It seems at least as doubtful that either of these can be substituted for a sequential strong undercutting $\geqslant_t$ in Principle I.

It is worth noting such possibilities, just because it emphasizes how difficult it would be to make moral judgments if we knew they arose. A tractable world (morally) would be one in which

a) Any action that is weakly undercut $\geqslant_t$ is strongly undercut $\geqslant_{t'}$ for some t' including t

b) There are no infinitely extended undercutting cycles

c) For every act A which is strongly undercut $\geqslant_t$ there is an act C which begins a strong undercutting $\geqslant_t$ of A

d) If C begins a strong undercutting $\geqslant_t$ of A then C begins a sequential strong undercutting $\geqslant_t$ of A.

Nicer yet would be one in which

e) If B undercuts$_0$ A then there is a t such that B begins a sequential strong undercutting $\geqslant_t$ of A.

But this seems too much to hope for.

I wish to leave as a question for further consideration whether the antecedent of Principle I can be weakened so as still to yield a valid principle. One would hope so, since the definition of "sequential strong undercutting $\geqslant_t$" uses very powerful machinery, which it would be better to avoid. (And it is not clear, for any action A such that $R_A > W_A$, how to rule out the theoretical possibility that there is an action C and a time interval t such that C begins a sequential strong undercutting $\geqslant_t$ of A.)

We should note that many problems that arise in the area of the subject of this section could be avoided if there were a time-discounting of the moral future. Questions about this issue, interesting in their own right, must be considered in a full development of the theory.

For the time being I shall suppose that no weakening of the antecedent of Principle I will do. (If this tentative supposition is false, what is said below can be modified so as to correspond to the new principle with weakened antecedent.) I suggest the following two principles:

1) If $R_A > W_A$ then [A is impermissible if and only if $(\exists B)$ $(\exists t)$ (B is an action available to the person & B begins a sequential strong undercutting $\geqslant_t$ of A)]

2) If $W_A > R_A$ then [A is permissible if and only if $(\exists C)$ (C is available to the person & A is part of C & $R_C > W_C$ & $\sim$ $(\exists B)$ $(\exists t)$ (B begins a sequential strong undercutting $\geqslant_t$ of C))][42]

1) replaces Principle VII) of Part IV. Let me close this section with a few remarks.

A. We have concluded that $R_A > W_A$ is not a sufficient condition for A's being permissible, and that $W_A > R_A$ is not a sufficient condition for A's being impermissible. So does any asymmetry remain? The asymmetry that remains is that $R_A > W_A$ can be shown to be insufficient for the permissibility of A by the finding of a suitable action B over the same time interval as A which is an alternative to A (where suitable means, among other things, that there is no act which begins a sequential strong undercutting $\geqslant_t$ of B); whereas $W_A > R_A$ is insufficient for the impermissibility of A only if A is part of a *longer* course of action meeting a certain description. Put roughly, if the moral future of an action were like its moral present, $W_A > R_A$ would be a sufficient condition for the impermissibility of A, whereas $R_A > W_A$ would not be a sufficient condition for the permissibility of A.

B. Our method of measuring the moral weight of a set of features was suggested by Principle VII) of Part IV. Does the substitution of 1) above for this principle change things so that the method proposed no longer works? It does not if, in our example to elicit a person's judgments about the permissibility of A, we are careful not to produce cases where $R_A > R_B$, $W_A > W_B$, $W_B << W_A$, *and* the person may believe that though consideration of just A and B makes A impermissible, considering longer courses of action changes things. I shall not enter here into the details of how the example to be put to the person should be constructed so as to avoid his having this belief. Using the terminology we have defined, we do not want to present him with examples where he will believe that though B undercuts$_0$ A, B does not begin a sequential strong undercutting $\geqslant_t$ of A. It should not be too difficult to avoid such examples.

C. Suppose $W_A > R_A$ yet $(\exists B)$ (A is part of B & $R_B > W_B$ & B is not impermissible (by 1)). According to what we have said, the act A is not impermissible. But the act A without doing the rest of B (or some such

[42] 2) is equivalent to: If $W_A > R_A$ then [A is permissible if and only if $(\exists$ C) (C is available to the person & A is part of C & $R_C > W_C$ & C is not ruled impermissible by 1))]

B) *is* impermissible, and can be shown to be so by our principles. Often, for short, in situations like this where the person does not do the rest of B (but only does the part which is A), we elliptically say that his act A was impermissible.

D. Finally, let us return to the questions with which this section opened. Can we say that "is part of a course of action which is such that ....." is a feature of an action A, and if so, how is the blank to be filled in? If the argument of this section is correct, in the case where it seems that $W_A > R_A$ yet A is morally permissible, the candidate for the extra R-feature of A is

F = "is part of a larger course of action C, where $R_C > W_C$, and there is no B and t such that B begins a sequential strong undercutting $\geqslant_t$ of C."

But F is not morally neutral since it refers to the R and W-lists, and this reference cannot be eliminated for reasons similar to those advanced in Part IV. Thus I conclude, once again here, that one cannot state everything in terms of the R and W-features of act A, and the introduction of higher order principles[43] is necessary in order to get things straight.

VII. Overriding, Outweighing, Neutralizing, and Related Notions

We have thus far gone along with the simple model's use of *one* inequality relation between sets of features of actions. But there are significant differences among the ways in which the presence of other features (or the obtaining of certain facts not represented by features of actions) can make an action morally permissible, even though the action has some W-features. An adequate model of the structure of the sort of moral view I am considering must in some way differentiate among some of these different ways, not all of which are happily classified under the rubric of "overriding" or "outweighing." In this section we shall list some of these ways, somewhat arbitrarily labelling them, and consider some of the problems they raise. We shall not, in this section, discuss how they are to be incorporated into an adequate model of the structure of a person's moral views. Thus, we are now concerned with setting problems and raising questions. The ways that I list are "pure" ways, and we shall not, at this point, be concerned with various combinations of them.

I. An act A, with W-features, prevents, avoids, etc., something bad, harmful, etc. It has no other R-features. If the act is morally permissible, let us

[43] In this case, principles 1) and 2) above.

say that its R-features *override* its W-features; if the act is morally impermissible, let us say that its W-features *outweigh* its R-features.[44]

It was to these notions (as well as to those in II below) that our previous discussion was especially meant to apply.

Relevant distinctions to make are whether

a) something like Principle VII) of Part IV is required

b) if the W-features of act A involve something bad to a person P, the performer of A is morally required to make either reparations to P, or amends to P, or to offer explanations to P, etc.[45]

I believe that in all cases where an act A, with $W_A \neq \phi$, avoids, prevents, etc. something worse (i.e., where its W-features are overridden), then something like Principle VII) is required. But even if this is correct, we cannot use Principle VII) to *explain* the notions of "overriding" and "outweighing," and this is not solely because the principle may legitimately apply to some other notions as well. The major obstacle to doing this is the following. For some of the ways we shall list in this section of a feature's making an action A morally permissible, even though $W_A \neq \phi$, some of the features which in these selected ways make A permissible do not belong on what we intuitively have in mind as the R-list. We want to first explain these various ways, and *then* to limit what can go on the R-list by excluding some features which play a role only via some of the selected ways. Since Principle VII) uses the notion of the R-list, we do not want to use it to *explain* one of the ways (viz., overriding and outweighing).

Are there actions where a) and b) do not go together? I believe that b) cannot be found without a); that is, whenever your doing A with non-null wrong-making features makes it incumbent upon you to explain your action to whoever is harmed by its W-features, or to make amends or reparations to these persons, then it is also the case that it would be impermissible to do the action if a suitable alternative (as defined by Principle VII) were available.

Can we have a) without b)? Can it be the case that you must do a suitable alternative act to A if one were available to you (and since one isn't, it's morally permissible to do A), yet even though A's W-features harm someone, you have no duty to make reparations or amends or to explain to them

[44] Here, as below, I ignore the issues raised in Parts IV and VI, and momentarily suppose that permissibility or impermissibility is determined by the way some general inequality goes. Such a simplification is useful as a first step in an intuitive explanation of the different ways.

[45] As in the case where you don't go to a dinner to which you promised to go, in order to minister to an accident victim whom you encounter on the way to the dinner. You must call your hostess "as soon as possible" (compatible with your best ministering), make explanations, etc.

why you've acted as you did? A plausible candidate for such a case of no duty to make amends or reparations would be good samaritan cases. In saving someone's life, I damage your property. I could not save that person without damaging your property or doing something at least as bad (though if I could have, it would have been wrong for me to damage your property). It may well be that I am not morally required to make amends or reparations to you.[46] However, even in this case one would think that I am required to give you some explanation of what's happened. In speaking of what I am required to do, I have put things too strongly. "Omitting to make reparations, amends, or explanations to those harmed by W-features (specifically listed) of act A" is a W-feature of a course of action, and functions just like other W-features. It can be overridden, etc.: it may be too dangerous to make explanations, too inconvenient given the slight harm caused, etc.

II. Act A, with $W_A \neq \phi$, though it does not prevent, avoid, etc., some harm, achieves some good. If (subject to the same qualifications as with I) act A is morally permissible, let us say that its R-features *overcome* its W-features; if A is morally impermissible its W-features *overshadow* its R-features. Here, as in I (if not more so) one feels that a) and b) of I obtain, and perhaps that b) in this case is stronger than it is in I. That is, that in this case there is a stronger obligation to make amends or reparations or offer explanations than in I,[47] and perhaps an obligation to make greater amends or reparations.

III. On some views, for some W-features (relations) T which take someone as a direct object, it is morally permissible to T someone who has T-ed you (or perhaps has only T-ed some other people), where the T of his act helped make it morally impermissible.[48] On such very contractual views, it might, for example, be permissible to steal from a thief, torture a torturer, etc.,

[46] Though perhaps the person saved is so required. This raises the question of whether in all such cases *someone* (though perhaps not the person who did the act) is morally required, if it is possible, to make amends or reparations to those harmed by overridden W-features of an act. Pursuing this issue leads one to issues similar to some discussed by welfare economists; see I. M. D. LITTLE, A CRITIQUE OF WELFARE ECONOMICS ch. VI (1960).

[47] When one reaches the point where one can make amends or reparations or offer explanations, some features override or overcome "not making amends or reparations and not offering explanations after a I situation" which do not override or overcome "not making amends or reparations and not offering explanations after a II situation," and the second outweighs or overshadows some R-features that the first does not.

[48] Take the first person who T-ed someone else. If his action is morally impermissible, he is from that time forth, a *T-person*. If his action was morally permissible, take the next person to T someone. The first one who does it impermissibly becomes the first T-person. (I ignore here the possibility that two persons simultaneously and impermissibly T each other.) The only way to become a T-person is by T-ing someone who is not a T-person, where the W-features of this act of T-ing are not overridden, overcome, etc. On the

without oneself becoming someone who is open to permissible thievery, torture, etc. Without worrying now whether it is a feature of the act or a fact about the situation which does so, let us say that in such situations of T-ing someone who T's, T is *neutralized*.

Certain retributivist views would hold, not only that sometimes it is permissible to T one who T's but that it is sometimes obligatory to T him, or to do some act with some other W-feature G. Thus on this view, it is wrong not to punish someone for certain offenses; and this apart from deterrent considerations; he just deserves it; and justice demands that he get it.[49] Again, without worrying over what it is that does so (is it a feature of an act that it's done to someone who has committed a wrong, or is it a fact about the situation?) let us say that in this case the W-feature of the act is *reversed*.

Sometimes T will be neither neutralized nor reversed, but it will be the case that T-ing someone who T's is less wrong, carries less W-weight than T-ing someone innocent of (wrongful) T-ing. Let us say in this case that T is *weakened*.

IV. Suppose T is on the W-list, and takes persons as direct objects. Suppose further that each of the persons who are the objects of T consent to being T-ed. It seems that sometimes this will have the consequence that T carries no moral W-weight, and sometimes it will have the consequence that T carries W-weight, but less than it would in the nonconsent case. In the first case, let us say that T is *dissolved;* in the second case let us say that T is *consent-weakened*.

V. Suppose that I have promised you that I will do an act A. You release me from this promise, and there are no third-party beneficiaries. Let us say, that in this sort of case, the feature "not keeping my promise" is *cancelled*.

VI. You extend yourself to do me a good turn. I am under an obligation to return it, if I can. Suppose you then intentionally go out of your way to (wrongfully) harm me. Let us say that in this case F = "omitting to

view we are considering, it is permissible to T a T-person; that is, the W-feature, T, of an action has no moral weight when the action is done to a T-person.

I do not wish to discuss here whether, for some feature T on the W-list, such a view is correct. It is a possible view, and an adequate account of the sort of moral structures we are discussing must be able to handle it.

[49] I shall not pause to consider the different views that fall under the retributivists' view that it's good that someone who has committed a wrong suffer; e.g., is it good that he suffer, or is it good that he suffer *because* he's committed the wrong (that his suffering be due to his having committed the wrong), or is it good that (his suffering be due to his having committed the wrong, and he know that his suffering is due to having committed the wrong)? And if more than one of these is held to be good, do they differ in degree of goodness?

reciprocate on unreciprocated good turn" is *destroyed* (from: The obligation is destroyed).

VII. I promise to meet you next week at a certain place and time, so that we can do something together. Before that time, I learn that you have died. Let us say that F = "involves omitting to keep my promise to meet you" is *nullified*.

VIII. You lied to me in order to get me to promise to do A. I believed your lie, and made the promise. Let us say that "involves not keeping my promise to do A" is *invalidated*.[50]

IX. Until now we have considered different ways in which the presence of features or facts either may lead to an action A's being morally permissible (even though A has features on the W-list), or may weaken the moral weight of some W-features. It may be that an adequate account of moral structure must also consider a quite different kind of relation. Perhaps some features or facts can *undermine* the operation (in one of the above-mentioned ways) of other features or facts upon W-features. Features or facts F operating in one of the ways discussed in I through VIII upon the W-feature of action A mark off exceptions to the rule that any act with all the features in W_A is morally impermissible. But perhaps there are also exceptions to these exceptions; that is, facts or features which, if present, prevent F from operating, as it normally does, in one of the ways upon W_A.

Consider, for example, Section 3.04 of the American Law Institute Model Penal Code (Proposed Official Draft, 1962), concerning justifiable use of force in self-protection. Put roughly, the structure of this section is as follows:

I. The use of force upon or towards another person is justifiable when P.

II. a) The use of force is not justifiable by I when

1) Q

or 2) R

except the R limitation upon the justifiable use of force under I does not apply if

a) S

or b) T

or c) U

b) The use of *deadly* force is justifiable under I only if V, and it is not justifiable (even if V) if

a) X

or b) Y

[50] V-VIII have only been sketched. The details needed to state them correctly would take us away from our major concern here.

except the Y limitation on the justifiable use of deadly force under I when V, does not apply if

a) Z

or b) M

Let w_1 = involves using force upon or towards another person; let w_2 = involves using deadly force upon or towards another person. Though it will often not be clear exactly why one structure is chosen rather than another, let us assume that we here have a four-levelled structure, and that there is good reason not to collapse it into fewer levels. We thus would have:

1) P stands in one of the ways to w_1
2) Q (R) *undermines* P's standing in one of the ways to w_1
3) S (T, U) *upsets* R's undermining of P standing in one of the ways to w_1
4) P & V stand in one of the ways to w_2
5) Q (R, X, Y) *undermines* P & V's standing in one of the ways to w_2
6) Z (M) *upsets* Y's undermining of P & V's standing in one of the ways to w_2.

Upsetting will be a 4-place relation (or more if one adds extra variables, e.g., one ranging over the ways). An important question is that of how many levels one is driven to; up to what n must an adequate theory use n-place relations of this sort? (I ignore the question of the possibility of reducing the number of places by formal gimmicks of various sorts.)

Up until now we have used just two-place relations (for the ways). Is there any strong reason for not continuing to do only this? Can't we incorporate all of the above information into the domain of the ways, and say, for example, that (where the "or" is the inclusive-or)

A) P & not-Q & (not-R or S or T or U) stands in one of the ways (or a combination of them) to w_1
B) P & not-Q & (not-R or S or T or U) & V & not-X & (not-Y or Z or M) stands in one of the ways (or a combination of them) to w_2.

This issue must be discussed in a full presentation of the theory. Here let us just note that many of the reasons that led us, in Part I, away from the deductive structure to one with open-ended lists of features, outweighings and overridings, etc., will apply here as well.[51] For reasons similar to the earlier ones we may want an open-ended list (for each feature or set of

[51] To be sure, the lawyers must write something down completely, without open-ended lists. (Though perhaps Section 3.02 of the Model Penal Code is meant to open the ends.) I should note that nothing I say here depends upon a claim that, given its content, Section 3.04 of the Model Penal Code cannot plausibly be reduced to a simpler structure.

features?) of what can undermine a feature or set of features' standing in one of the ways to W-features, and open-ended lists of what can upset such underminings. If so, this would prevent us from working just with things like A) and B) above. Obviously, if we take the course of not working only with things like A) and B) above, we have the pressing problem of describing the more complicated structure which is to substitute for the one which uses only the two-place relations.

Many other two-place notions in addition to those in I-VIII above, might be put forth; e.g., we might say that a W-feature is *precluded* in cases where, e.g., it is impossible to keep a promise (can't implies not-ought cases). My purpose here is not to proliferate notions for its own sake, but to raise issues which require further consideration. In addition to the ones mentioned, our brief discussion in this section raises several other serious issues. I shall do little more than list them here.

A. What are the various ways in which the presence of some features of an action A may, *ceteris paribus,* prevent the action from being morally impermissible, even though $W_A \neq \phi$? How are these different ways to be distinguished and embedded within a general theory of moral structure? Which of the ways S_1 (viewed as relations in which things can stand in to W-features) must have only R-features in the S_1-image of W? For example, one would not expect the relation in V and VII above to have only R-features in their image of W. What principles can one formulate which govern each of the ways? Which ways require special principles, and which principles? How will the theory handle combinations of ways, and what special problems does this raise?

B. How are "acts of omission" to be handled within the theory (consistently with our discussion of features in Part III)?

C. For features (or facts or conditions) which, according to III - VIII above neutralize, reverse, cancel, dissolve, destroy, nullify, invalidate W-features, should the denial of these features (or facts or conditions) be built into the W-features themselves? I shall not attempt to list here the plethora of considerations (in addition to the ones related to those discussed under IX above) relevant to this question.

D. We have spoken above of some features (or facts) weakening the moral weight of W-features. There seem to be at least two ways in which one might try to handle this.

1) Each W-feature, as well as each set of W-features, always has one moral weight. What looks like weakening (or the lessening of the weight) is just the result of something's being put on "the other side of the scales." The "moral scales" come to rest where

they do, not because some weight is lessened but because some extra weight is put on the other side.

But this view would have the consequence that it is the one unique full weight of the W-feature which must be used in something like Principle VII of Part IV, and which also determines what amends, reparations, etc. should be made. This consequence would certainly be implausible for the cases which someone holds fall under weakening in III above. For if, according to such a person, some W-feature w_1 of some act B is weakened, one should *not,* in determining whether some alternative action has less weighty W-features (under Principle VII) than action B has, treat w_1 of B as having its full weight. For there might be an alternative action A such that the full weight of w_1 is greater than that of W_A while the weakened weight of w_1 had by B is less than that of W_A. Thus it seems that the method of explaining and handling weakening put forth by 1) has undesirable consequences when combined with some principle like Principle VII of Part IV.

2) The alternative possible way to handle weakening is to admit that the weights of some features are really lessened (and not just partially compensated for while remaining the same, as in 1)).

The obvious questions are: How is this possible way to be specified in detail, and how must the structure be modified to accommodate it? How does this affect the program of measurement set forth in Part V? It seems that one could, by determining a person's judgments only for cases in which weakening does not operate, obtain the same measurement results as before. Can a similar measurement procedure be devised to measure the moral weights of *weakened* features in specific situations? Do we also need a notion of the strengthening of a W-feature, of its weight being increased, and if so, how is this to be incorporated within a systematic theory of moral structure?

E. In Parts II - VI we have considered only other features or sets of features overriding or overcoming, etc., W-features. Must one consider things other than features of actions in order to account for a person's judgments of moral impermissibility, e.g., facts about the situation which are not happily incorporated into features of actions? If so, what apparatus is needed to handle this extension?

F. We have thus far avoided discussing the issue of whether some of the features on the list must incorporate a person's beliefs. That is, we have neglected the issues that have led into the morass of discussions of subjectively right (ought) and objectively right (ought). How are such issues to be handled within the sort of structure we are discussing? Can some of the distinctions made in I - IX of this section show a way through these issues; i.e., might belief be appropriately included for some of the ways (and under-

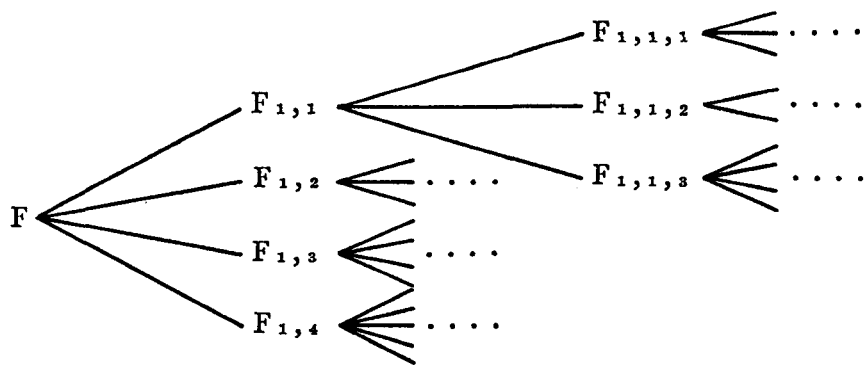
minings, upsettings) and not for others? If so, this might be relevant to some other questions mentioned earlier; e.g., to whether the denial of neutralizing, cancelling, reversing, dissolving, destroying, nullifying, invalidating features or facts should be built into the W-feature.

VIII. Further Issues

It is unnecessary for me to restate, to the reader who has come this far, that this paper is an exploratory study meant to raise further issues for investigation as well as to propose tentative solutions to certain problems. Now I wish to indicate some questions and issues, in addition to those left open in previous sections, which require further study.

1) What is to be said of the possibility that always or sometimes there's not *one* feature of an act such that if an act has it, it is, *ceteris paribus,* morally impermissible, but rather that a conjunction of features is like this? In this case, should the W-list consist of conjunctions of features (with some different entries on the lists having common conjuncts) or is there some simpler apparatus to handle this? Similar questions, obviously, can be asked about the R-list.

2) Given the kind of unity and coherence people's moral views have, there is some reason to want some of the entries on the lists not to be merely features (or conjunctions of features), but rather branching tree structures of features (or their conjunctions). Thus,



Any action which has feature F has each feature referred to by the expressions obtainable by (repeated) deletion of numerals at the end of the subscript of F. Thus any action with, e.g., feature $F_{12,6,9,4,7}$ has features $F_{12,6,9,4}$, $F_{12,6,9}$, $F_{12,6}$ and F_{12}. Intuitively, $F_{12,6}$ is a way of realizing F_{12}, etc.

Let us call the set of all features to the right of a feature F on a tree which can be reached by following a path from node F and going to the right, the

descendant set of F. One might put forth the following principle of inference within a moral system:

$$\frac{S_1 > S_2}{S_3 > S_4},$$

where S_3 is a non-null subset of the union of the descendant sets of the members of S_1, and S_4 is a non-null subset of the union of the descendants of the members of S_2. But reasons similar to those which led to rejection of the deductive structure in Part I and to the consideration of the simple model in Part II, may make such a principle of inference too strong. If so, and tree structures of features are appropriate ways of representing the unity and coherence of a person's judgments, what rules are to govern them?

3) Given open-ended lists of features (and tree structures of features?), some of which contain variables (e.g., "leads to the death of n-persons"), is the number of entries on each list finite or infinite? Would a correct and complete moral view require finite or infinite lists? (Can there be a correct and complete moral view?) If infinite, are the lists recursive? recursively enumerable? are complete and correct moral systems finitely axiomatizable? Something must be said about what kinds of reduced lists are permissible. What criterion can be formulated which would allow "leads to the death of n-persons" on the W-list (and wouldn't require that, for each number m, its instantiation for m appear on the W-list) yet would not have just one utilitarian feature on the list even if the person was a Sidgwick-type utilitarian and utilitarianism were true?

4) What is one to make of talk of open-ended lists of features? For at any given time, presumably only finite specific lists are needed to account for the judgments the person has actually made previous to that time. The openness of lists was meant to mark off a problem. It seems that given any finite list that we or he could construct, there is a certain situation which if described to the person, would lead him to say that some feature not on either presented list was a morally relevant feature of the situation: "Oh yes, I overlooked that one." Though this feature wasn't necessary to account for any of his previous judgments, in some way the description of his view must "include" the feature, must take account of the fact that he *would* accept the feature as morally relevant, and he would take it into account in arriving at a judgment about a situation which exhibited it. It is not clear what well-understood formal device best achieves the legitimate purposes of the vague talk about open-ended lists. (For some persons, nothing like open-ended lists will be appropriate, and indeed some sort of closure condition will be required; e.g., "nothing belongs on the W-list except $F_1, \ldots, F_n$.")

5) How is the structure to handle second-order (and higher?) features; e.g., the W-features: persuading someone to do something impermissible, forcing someone to do something impermissible, offering someone something to do an impermissible act, glorifying, praising, rewarding impermissible acts, punishing, ridiculing required acts, leading someone to abandon correct moral principles? Sometimes the particular first-order features needn't be specified, or specified precisely, to account for the person's judgments; sometimes they can't be (as when one person persuades another to do some unspecified impermissible act), and sometimes it will be important to know what exactly the first-order feature is. How should all this be represented within a structure?

6) What sorts of structural conditions using the notion of necessity are appropriate? Would it be appropriate to require that if necessarily $-(F_x \equiv G_x)$ then F and G override, are overridden by, exactly the same sets of features, etc.?

7) Most persons apply different moral standards to different kinds of beings, e.g., children and adults. Some persons may do so for different groups of adults, the distinctions depending upon, e.g., social class, religion, race. Should this be represented (ignoring the additional ways discussed in Part VII) by *one* pair of lists (one member of the pair being a W-list, and the other an R-list) where the lists contain "conditional-features," or by *separate* pairs of lists for each group for which there are distinct moral standards?

8) One should define various notions of a complete moral system (one yielding for every possible case of a certain sort a judgment that the act is morally permissible or morally impermissible). Different notions will be defined for different kinds of information fed into it. For example, one notion would be that, for given lists of R and W-features, for any two arbitrary subsets of R and W, the system yields an inequality between them. (See Condition 2 of Part III.) Further notions would be defined for risk, and uncertainty situations. Are moral dilemmas or drawing-the-line problems reasons to believe that there does not "exist" (everyone grants that we don't know one) a complete and correct moral system?

9) One wants some way of describing ways in which the members of a list can cohere. Various writings on the coherence theory of truth would be suggestive here. One might define a W-features' being a *taboo* as its not cohering in certain specified ways with some other members of the W-list.

10) The structure described is concerned with the relatively gross (though difficult enough!) distinction between moral permissibility and moral impermissibility. Various extensions immediately suggest themselves, e.g., to *degrees* of moral impermissibility; and to other finer distinctions.[52] How must

[52] Distinctions analogous to those discussed, for example, in R. Chisholm, *Supereroga-*

the structure be supplemented to account for these more refined judgments?

11) We want some way to represent a person's belief that nothing, or something, or nothing or something of a certain sort overrides, outweighs, etc., some set of features, and to incorporate such beliefs within the general structure, and have them be one of the things the structure accounts for. Various obvious ways suggest themselves. (If *nothing* overrides a set of features, this set must get special treatment in the measuring process of Part V.)

12) We have thus far considered problems stemming from a person's lack of confidence in the correctness of stated exceptionless moral principles due to the wide variety of cases which are possible or may arise. He can't anticipate all possible cases, etc. But a person may also have varying degrees of confidence about a specific feature's being on one of the lists or about a specific outweighing (which he may strongly, tentatively, or with some hesitation accept). We want some systematic way to make sense of different degrees of confidence, to measure them, and we want to formulate principles specifying the role they play in generating a person's moral judgments.[53]

13) What, if any, are the interesting ways in which two particular moral views having the same general structure can differ if they generate exactly the same moral judgments about all possible particular acts?

14) May *indexical* expressions or proper names be built into features on the lists; e.g., "harms *my* wife," "involves eating on Yom Kippur"? What constraints are there on how they may appear, and on what else must appear if they appear?

15) How are double-effect type issues to be handled within the structure? Must intentions be introduced within the features?

16) One is tempted to use the structure as a weapon in discussion. (R. M. Hare's use of a similar weapon in his book *Freedom and Reason* is too strong.) One is tempted to say that if a person judges that act A is morally impermissible and act B is morally permissible, then it's incumbent on him to produce a feature on one of the lists had by one of the actions and not another. The argument for this would be that *you* cannot go completely through the two open-ended lists showing him that A and B have exactly the same list features, but *he* can produce the feature which distinguishes A and B. But if he produces a feature which is had by A and not by B, and you deny that this feature is on either list, is it incumbent upon him to produce an argument for this feature's being on one of the lists? Or is it incumbent upon you to produce an argument for the feature's not being

tion and Offense: A Conceptual Scheme for Ethics, 5 RATIO 1-14 (1963), and R. Chisholm and E. Sosa, *Intrinsic Preferability and the Problem of Supererogation,* 16 SYNTHESE 321-31 (1966).

[53] Suggestive material can be found in I. Levi's GAMBLING WITH TRUTH ch. VIII.

on either list? Difficult questions arise about where, if anywhere, burdens of argument or proof lie. And how are we to explain: "it is incumbent upon him to"? Is it just that he be able to (in how long a time?), or that not doing it when requested is a W-feature (which can be overridden; e.g., he has more important things to do)? And what follows from a person's being unable to discharge his burden; that one of his judgments is mistaken (surely this doesn't follow), that one of them is unsupported, that he should shut up, that he's not serious?

17) Condition VII of Part IV, though growing out of the earlier discussion, was formulated without any requirement of the relevance of the alternative act B to the R-features of A. Thus, someone might object that condition VII owes its plausibility to the assumption that B is a substitute for A, and achieves the same or similar goals, and if this is not so, then VII is too strong. It is certainly worth investigating the systematic consequences of adding to condition VII various precisely formulated requirements concerning the relevance or similarity of R_B to R_A.

18) Some people's views may exhibit different metarules. To take two simple examples:

A. *Permissive Rule:* If the system of inequalities and principles does not yield the result that an action whose only morally relevant features are all and only those in set S is morally impermissible, then any action with just those morally relevant features is permissible.

B. *Strict Rule:* If the system of inequalities and principles does not yield the result that an action whose only morally relevant features are all and only those in set S is morally permissible, then any action with just those morally relevant features is impermissible.

(Of course it may be that a person's views exhibit no such rule.) One might attempt to represent a person's views which might be interpreted as exhibiting one of these metarules as containing some rule of inference yielding inequalities which complete his incomplete set of inequalities. What reasons might be advanced for preferring the metarule representation which does not involve adding first-level inequalities, to the rule of inference representation which does?

19) Can one incorporate some legitimate device which functions to produce self-reference, and arrive at an action A which can be interpreted as the act of judging that A is morally impermissible? If so, one will get the result that either there is some morally impermissible act which it is morally

impermissible to judge (not just *say*) is morally impermissible, or there is some morally permissible act which it is morally permissible to judge is morally impermissible.

20) Much work needs to be done on the subject of the "logic" of presumptions, accounts of what it means to say that presumptions can be overridden only by *special* reasons, etc. The relevance to the sort of structure we have been describing is obvious.

21) Until now, I have spoken freely of the model's accounting for a person's judgments about the moral permissibility or impermissibility of particular actions. For various reasons the model cannot do this, and requires supplementation.

Consider the following inference.[54]

1) Act A has features $F_1, \ldots, F_n$.
2) Each of these features is on one of the lists.
3) The subset of the F_i on the W-list $>$
   the subset of the F_i on the R-list.
4) A has no other features, in addition to $F_1, \ldots, F_n$, on either list.

$\therefore$ 5) A is morally impermissible.

Suppose that a person knows that 1), and suppose further that 2) and 3) are true of the particular structure correctly ascribed to him. Suppose that this inference pattern is built into the structure of his views, in that if he also knows 4) then he makes, or would make, the particular judgment 5). [Of course this particular inference pattern won't be built into the structure of his views since it ignores the complications about larger courses of action and alternative actions.] But the person may not know that 4). It may be that

a) The person believes, but does not know, that 4).
b) The person knows that A has some other features on the lists though he does not know which features.
c) The person believes, though he does not know, that A has some other features on the lists, and he does not believe, for some other particular features on the lists, that A has these.
d) The person does not believe that A has some other features on the lists, nor does he believe that A has no other features on the lists.

It seems to me likely that a person in situation a) will judge that A is morally impermissible (though if he has little confidence in his belief, perhaps he will not). Situations b) and c), of course, encompass many interesting cases. Looking just at c):

[54] I ignore the complications introduced by Part IV about alternative courses of action, and also ignore the complications introduced in Part VI about longer courses of action. These extra complications just reinforce the point to be made here, and would unduly complicate my exposition.

The person believes that A has some other features on the lists, though he does not believe, of some specific other features on the lists, that A has these.

We might have, in addition, that the person believes any one of the following:

Each of the other features is on W.

Each of the other features is on R.

Some are on W and some are on R.

Some are on W (and he has no beliefs about whether others are on R).

Some are on R (and he has no beliefs about whether others are on W).

Or we might have that the person knows that *if* A does have some other features on the lists then these features are such that ..... (with various possible fillings in of the blank). And for each of the cases, he may have beliefs about some of these (unspecified) properties outweighing or overriding others, about how the inequality goes when some are conjoined with the ones he specifically knows of, etc.

It seems clear that for situations b) - d), some further apparatus must be conjoined with the structure we have been discussing, in order to yield a person's particular judgments about particular actions. The details of this apparatus may vary from person to person; people may differ in how they make judgments in situations b) - d). A similar argument can be offered for simplified inferences yielding the judgment that an action is morally permissible. When the complications about alternative actions and larger courses of action are taken into account, it becomes even more clear that, in order to account for a person's particular moral judgments, the structure we have been discussing must be conjoined with a further apparatus about how the person arrives at beliefs in, and what he assumes with what confidence as a basis of inference in, situations of incomplete knowledge.

Secondly, we must further delimit the particular judgments to be accounted for by the structure we have been discussing. One would not expect the structure to play a major role in accounting for a person's belief that a particular action A is morally impermissible, which belief the person holds because someone he trusts told him that A is morally impermissible.[55]

Thirdly, one needs a distinction similar to that which linguists make between linguistic competence and linguistic performance.[56] The actual judgments which a person makes will depend upon various limitations common to

[55] Persons who hold that if someone is to have made a moral judgment that an action is impermissible, then he must have certain kinds of moral reasons in *support* of the judgment, must deny that in this case the person has made a moral judgment.

[56] See N. CHOMSKY, ASPECTS OF THE THEORY OF SYNTAX ch. I, sec. 1 and 2; and for some critical remarks, G. Harman, *Psychological Aspects of the Theory of Syntax,* 64 JOURNAL OF PHILOSOPHY 75 (1967).

all persons, and upon some special to him (or at any rate, not common to all); e.g., limitations of attention span, of memory, limitations on the complexity of information which can be manipulated and processed, limitations on the amount of time he is willing to spend thinking about moral problems. The actual judgments he makes will also depend upon various exogenous factors; e.g., whether he has a headache, whether there's noise which prevents him from thinking as clearly as he otherwise would, whether he's interrupted while thinking and loses some thought which he doesn't later remember. We may think of the structure we have been discussing either as a model of an idealized moral judge in idealized circumstances (i.e., ignoring the various limitations and exogenous factors), or as *one component* of an adequate psychological model of the person.

22) One needs a discussion of what it means to ascribe such a structure to a person, of what it means to say he "internalizes" such a structure. It doesn't mean that he always arrives at his moral judgments by explicitly and consciously referring to such a structure. I do, however, want the assumption to be based upon something more than the claim that the hypothesis that there is such a structure somehow realized inside him most simply and elegantly accounts for the moral judgments he makes. It would be upsetting, in this case, if the structure could not account (when combined with further theory) for his conscious reasoning in arriving at moral judgments, and the sort of considerations he adduces (and the way in which this is done) in support of his moral judgments. Put vaguely, the structure shouldn't be foreign to the way he actually reasons about moral matters, and he shouldn't find it, when presented to him, foreign.

23) The obvious ways in which the sort of structure we have been discussing might be found defective is that it can't account for some of the person's judgments about moral impermissibility (and this is true, for every person), or it "accounts" for more than he would make, or a simpler alternative structure accounts for his judgments about moral impermissibility in a neater and more elegant way.

It is worth mentioning two other ways in which the sort of structure we discuss might be found to be inappropriate (and these ways qualify what is said in the previous sentence). The structure is designed to play a role in accounting for only some of a person's moral judgments; viz., those about the permissibility and impermissibility of actions. It does not treat of judgments about goodness, virtues, ideals, responsibility, etc. It might turn out that the simplest total structure which accounts for *all* of a person's moral judgments which has the structure we have been discussing as a part, is less simple, elegant, or adequate than an alternative structure which accounts for all of

a person's moral judgments and which does not contain our structure as a part. Secondly, as I said at the very beginning of this essay, I do not claim that *everyone's* views about the moral impermissibility of actions exhibit the sort of structure we have been discussing. Some other kind of structure might account for a wider range of views about the moral impermissibility of actions while including the ones we have had in mind (e.g., it might also account for the views of persons in other cultures, or the views of all or of more persons in our culture), and this might lead one to reject the sort of structure which has been our subject.[57]

[57] One year after the completion of this essay, I am led to add a gnomic footnote, which I hope to explain elsewhere. It now (September, 1968) seems to me that if one were completely successful in carrying through the program of this paper (which is *far* from having been done here), one would have produced a *Tractatus Logico-Ethicus*. What is needed, perhaps, is an Ethical Investigations.

BIBLIOGRAPHY

This bibliography contains complete references to the books cited in this essay.

The American Law Institute, *Model Penal Code: Proposed Official Draft,* May 1962

Butler, R. J., ed., *Analytical Philosophy, Second Series,* Basil Blackwell, Oxford, 1965

Chomsky, Noam, *Aspects of the Theory of Syntax,* The M.I.T. Press, Cambridge, Mass., 1965

Colodny, R. G., ed., *Mind & Cosmos,* University of Pittsburgh Press, Pittsburgh, Pa., 1966

Goodman, Nelson, *Fact, Fiction, and Forecast,* Harvard University Press, Cambridge, Mass., 1955

Hare, R. M., *Freedom and Reason,* Oxford at the Clarendon Press, 1963

Hart, H. L. A. and A. M. Honoré, *Causation in the Law,* Oxford at the Clarendon Press, 1959

Hempel, Carl G., *Aspects of Scientific Explanation,* The Free Press, New York, 1965

Levi, Isaac, *Gambling with Truth,* Alfred A. Knopf, New York, 1967

Little, I.M.D., *A Critique of Welfare Economics,* Second Edition, Oxford University Press, Oxford, 1960

Luce, R. D., R. Bush, and E. Galanter, eds., *Handbook of Mathematical Psychology,* Vol. I, John Wiley and Sons, New York, 1963

Machol, R. E., ed., *Information and Decision Processes,* McGraw-Hill, New York, 1960

Ross, W. D., *Foundations of Ethics,* Oxford at the Clarendon Press, 1939

Ross, W. D., *The Right and the Good,* Oxford at the Clarendon Press, 1930

Sidgwick, Henry, *The Methods of Ethics,* Seventh Edition, Macmillan and Company Ltd., London, 1907

Suppes, Patrick, *Introduction to Logic,* D. Van Nostrand Co. Inc., Princeton, New Jersey, 1957

Thrall, R. M., C. H. Coombs, and R. L. Davis, eds., *Decision Processes,* John Wiley and Sons, New York, 1954

Von Neumann, J. and Oscar Morgenstern, *Theory of Games and Economic Behavior,* Princeton University Press, Princeton, New Jersey, Third Edition, 1953