1985

Introduction from *Morality and Universality: Essays on Ethical Universalizability*

Nelson T. Potter
University of Nebraska - Lincoln, kjohnson6@unl.edu

Mark Timmons
University of Illinois

Potter & Timmons in Nelson Potter and Mark Timmons (editors), *Morality and Universality*, ix-xxxii.

INTRODUCTION

In the past 25 years or so, the issue of ethical universalizability has figured prominently in theoretical as well as practical ethics. The term, 'universalizability' used in connection with ethical considerations, was apparently first introduced in the mid-1950s by R. M. Hare to refer to what he characterized as a logical thesis about certain sorts of evaluative sentences (Hare, 1955). The term has since been used to cover a broad variety of ethical considerations including those associated with the ideas of impartiality, consistency, justice, equality, and reversibility as well as those raised in the familar questions: 'What if everyone did that?' and 'How would you like it if someone did that to you?'

But this recent effloresence of the use of the term 'universalizability' is something that has deep historical roots, and has been central in various forms to the thinking about morality of some of the greatest and most influential philosophers in the western tradition. While the term is relatively new, the ideas it is now used to express have a long history. Most of these ideas and questions have been or can be formulated into a principle to be discussed, criticized, or defended. As we discuss these ideas below this principle will be stated on a separate numbered line.

The concepts of justice and equality were closely linked in Greek thought. These connections between these two concepts are apparent even in two authors who were hostile to the connection, Plato and Aristotle. They attempted to defuse the connection by emphasizing that justice or equality requires treating unequals unequally (see Vlastos, 1973). Aristotle writes:

> Now if the unjust is unequal, the just must be equal; and that is, in fact, what everyone believes without argument ... the just involves at least four terms: there are two persons in whose eyes it is just, and the shares which are just are two. Also, there will be the same equality between the persons and the shares; the ratio between the shares will be the same as that between the persons. If the persons are not equal, their [just I shares will not be equal; but this is the source of quarrels and recriminations, when equals have and are awarded unequal shares or unequals equal shares.[1]

This discussion could be condensed into the following principle.

(1) Treat equals equally, unequals unequally.

Jesus in the *New Testament* gives an influential statement of what has long been called the "Golden Rule," but which is more often discussed by philosophers under the title of a moral principle of reversibility:

(2) Do unto others as you would have others do unto you.

There seem to be problems with the Golden Rule when we attempt to use it to determine what our obligations in specific circumstances are. It is because of these difficulties that, as is remarked below, philosophers, when they get their hands on this principle tend to turn it into a non-substantive principle (see Singer, 1963; Gewirth, 1978). Another course is followed by Narveson in the lead essay of this anthology, who considers its claims as a moral principle in a variety of different formulations and rejects them all. Apart from such specific difficulties, part of the basic insight of the Golden Rule remains alive in those philosophers such as Kurt Baier who make significant use of such related notions as reversibility or reciprocity.

One of the most influential of moral philosophers ever, Immanuel Kant, proposed several different formulations of what he called "the categorical imperative." And he called the categorical imperative the "supreme principle of morality." This latter phrase apparently means that he believed that this principle was the only *moral* principle needed to determine our moral obligations in any particular circumstance (i.e., with reference to any given maxim). Other statements used in arriving at such conclusions about our moral obligations would be one and all non-moral statements. According to this view, any other correct moral principles or rules, such as "Don't tell lies," or "Be beneficent," would be a subsidiary principle or rule arrived at using only a single moral principle, the categorical imperative itself. What is often called the "first formulation" of the categorical imperative is a version of what we will shortly be referring to as a substantive universalizability principle. Here are two different statements of it:

(3) Act only on that maxim through which you can at the same time will that it should become a universal law.
(4) Act as if the maxim of your action were to become through your will a universal law of nature (Kant, 1785, AK., IV, 421).

The differences between these two versions are important for its application, but for our present purposes we may refer to either statement indifferently as versions of what we will call, following the usual

practice, the "first formulation" of the categorical imperative. Both versions contain the key phrase "universal law" which makes it clear that Kant's attempt is to use a substantive universalizability principle as the basis of his entire moral theory. This is the most ambitious claim ever made for a principle of ethical universalizability.

Henry Sidgwick, writing about a century ago, kept a notion of universalizability alive within the utilitarian tradition:

(5) [I]f a kind of conduct that is right (or wrong) for me is not right (or wrong) for someone else, it must be on the ground of some difference between the two cases, other than the fact that I and he are different persons (Sidgwick, 1901, p. 379).

Another formulation which seems likely to yield quite similar results to Sidgwick's principle is one that has been mentioned by more than one recent writer; here it is given in a formulation from a recent book by an author who is a contributor to the present volume, Wlodzimierz Rabinowicz:

(6) Moral properties of things (persons, actions, states of affairs, situations) are essentially independent of their purely 'individual' or 'numerical' aspects (Rabinowicz, 1979, p. 11).

Another direction has been taken by another much discussed writer who also contributes to this volume, Marcus Singer:

(7) If the consequences of everyone's doing some action *x* would be undesirable, then no one ought to do *x* (Singer, 1961, p. 66).

In another recent influential essay Bernard Williams imagines a world in which we have the power to control, alter, and improve all of people's personal qualities to the point where each is brought up to the level of all the others in qualifications. He writes,

> In these circumstances, where everything about a person is controllable, equality of opportunity and absolute equality seem to coincide; and this itself illustrates something about the notion of equality of opportunity (Williams, 1962, pp. 128-9).

The idea of absolute equality appears to be one of the more radical theses that a concept of universalizability could yield: the idea that all persons should be treated in exactly the same way period.[2] Such an

idea has had few if any defenders, but as the present quotation from Williams indicates, its existence as a possible ideal may be used to help clarify by contrast other related conceptions. Williams' thought may be formulated in the following principle:

(8) Where all differences between persons are controllable, equality of opportunity and absolute equality coincide.

The connections with universalizability are also clear in the statement of the first part of John Rawls' principle of justice:

(9) [E]ach person is to have an equal right to the most extensive basic liberty compatible with a similar liberty for others (Rawls, 1971, p. 60).

The connections with universalizability are to be found in the words "equal" and "similar."

The above nine principles aim to represent both the history and the possible scope of concepts of ethical universalizability. There are some who think that it is not a single concept that has such a scope, but at best a family of concepts that exhibit family resemblance, and at worst a miscellaneous assortment of ideas arising from various puns on key words such as "equal" and "universal." And we wish to regard these questions about the unity of the concept of universalizability and about the importance or soundness of the concept in its applications (the Nakhnikian issue) as open ones.

Singer, though not exactly a "sceptic" in either of the senses just specified, remarks in his paper below that not only has 'universalizability' been used to cover a variety of distinct theses, but '... the ... term has been so promiscuously generalized as to cover a variety of only tenuously related matters.'

These points raise questions about the connections, however tenuous, among these principles, which in turn raise more fundamental questions about their meaning, justification, and application. The papers contained in this anthology represent some of the most recent thinking about these fundamental questions regarding universalizability principles and related matters. In editing this volume, we have sought to arrange the papers according to what seems to us as a helpful way of distinguishing among types of universalizability principles. In what follows, we shall introduce our scheme of organization and explain how the papers fit into this scheme.

We have found it convenient to classify universalizability principles of the sort listed above according to whether they are *non-substantive* or *substantive*. What we mean in calling a principle 'non-substantive' is, roughly, that such a principle does not entail, either alone or together with other non-moral premises, any moral conclusions of the sort that something (some action, person, state of affairs) has a certain moral property. Rather, it is only in connection with a moral Judgment or statement to the effect that a thing of a certain sort has a certain moral property that a non-substantive principle can be used to derive any moral conclusions. By contrast, a 'substantive' ethical principle is one which, either alone or together with other non-moral premises, can be used to derive moral conclusions.

This distinction between non-substantive and substantive moral principles is the basis for what seems to be a main division between two types of universalizability principles. Thus, the so-called 'principle of universalizability,' alternatively expressed by the first two principles on the above list, is typically interpreted as a principle of ethical consistency, which as one author has put it: 'may be seen as a purely hypothetical thesis. According to it, a certain moral claim applies to an object *only if* similar moral claims apply to similar moral objects' (Rabinowicz, 1979, p. 14). The idea is that in order to derive a moral claim about (for example) the moral quality of some action in a certain situation, this principle must be taken in conjunction with a moral claim about the possession of the same moral quality by a similar action in a similar situation. By contrast, substantive universalizability principles, such as Kant's categorical imperative and Marcus Singer's generalization argument, are not hypothetical in this sense; rather they set forth a standard or test for determining in connection with other non-moral information, the moral acceptability of something. In the case of the categorical imperative, those actions whose maxims cannot be willed consistently as universal laws of nature are unacceptable. The generalization argument on the other hand (If the consequences of everyone's doing some action x would be undesirable then no one ought to do x without a reason) sets forth a standard forbidding those actions the general doing of which would produce undesirable consequences.

Reflection on important differences between the principles of Kant and Singer in particular, and on the differences between two of the main normative ethical traditions in general, viz., the Kantian and utilitarian traditions, suggests a further subdivision within the class of substantive universalizability principles. Thus, if we use the term, 'consequentialist', in a general way to denote not only utilitarian prin-

ciples, but also those associated, though strictly speaking non-utilitarian, principles,[3] then we can distinguish between Kantian and consequentialist universalizability principles.

Of course, this scheme for dividing universalizability principles is only intended to provide a rough means for classifying such principles, one which though useful, does not seem to make mutually exclusive divisions. For example, Jonathan Harrison, in his contribution to this volume, seems to be proposing a consequentialist moral theory, though one incorporating considerations of Kantian universalizability. Moreover, some principles are not clearly classifiable as either substantive or non-substantive, e.g., the Golden Rule. In its most common formulation 'Do unto others as you would have them do unto you,' it appears to set forth a test or criterion for morally acceptable action. Gewirth has put it this way. 'This criterion consists in the agent's desires or wishes for himself qua recipient: what determines the moral rightness of a transaction initiated or controlled by some person is whether he would himself want to undergo such a transaction at the hands of another' (Gewirth, 1978, p. 133).

Now the problems with the Golden Rule so interpreted have been recorded many times, going back at least as far as Kant (see Kant, 1785, AK, IV, 430n). In attempting to preserve the spirit of this principle, if not the letter, many philosophers have sought to find an acceptable formulation of the rule. And what is interesting is that some, though not all, of these attempts have in effect transformed it from an apparently substantive principle into a non-substantive principle (see Singer, 1963, and Blackstone, 1965).

With these points in mind, let us turn to some of the specific issues that have been raised in connection with each of these types of universalizability principles and briefly indicate how the papers in this collection are related to such issues as well as to each other.

ETHICAL UNIVERSALIZABILITY: A VARIETY OF THESES

We placed Jan Narveson's paper at the head of the anthology because it provides a broad survey of some of the main theses that have been discussed or might be discussed under the title of 'universalizability.' This survey to some extent cuts across the distinctions embodied in the arrangement of the rest of the essays in the present volume, so that there is scarcely another single appropriate place for it in our table of contents. It also provides an excellent introduction to the present state of discussions of universalizability. The range of Narveson's

discussion is great, emphasizing a point we have already made in this introduction, that the term 'universalizability' is today used to cover quite a large family of ideas. Some of these theses when they are considered as the only contribution that the concept of universalizability has to make to morality, make the contribution of that concept rather peripheral. Others, when considered by themselves, make universalizability to be close to the essence of morality. After taking us on a tour of some of the more peripheral conceptions, Narveson zeroes in on some of the more important conceptions, the ones in which he finds himself the most interested.

In the course of his discussion he criticizes certain formulations of the Golden Rule and other reversibility criteria as causing problems seemingly no matter how they are formulated. He does not distinguish sharply between U-formulations that are utilitarian or consequentialist and those that are not. In fact it is an interesting surprise to find this well-known defender and explicator of utilitarianism to be indicating some doubts about certain aspects of utilitarianism (see especially note 31) and to be defending the view that universalizability is close to the essence of morality. His final formulation is:

U16 *R* is an acceptable moral rule only if there is sufficient reason in terms of their own values, for all moral agents who have reason to decline the ruleless state to accept *R*.

The qualification referring to those who "have reason to decline the ruleness state," would omit or exclude only those who prefer or at least have no reason on balance to reject the Hobbesian state of nature. Fortunately such a person would be both rare and unusual. Apart from this exclusion, which we have every reason to expect is insignificant, the point is that there are certain rules which everyone will have an interest in everyone's following. Such a skeleton of rules for social relations, if they are fairly well in place, can give us assurance in dealing with others. Many of these rules, as we may expect, would be negative rules of non-interference. And since U16 refers to "all moral agents," it puts aside golden rule issues about whether agent or patient preference should be followed.

U16 is then finally a principle that is not merely proposed abstractly for our decision-making; it is instead based on a conception of what morality is for (regulating the skeleton of all personal relations so that individuals can within this structure accomplish their own goals). And

it places one specific concept of universalizability at quite a central location in the structure of morality.

UNIVERSALIZABILITY AND ETHICAL CONSISTENCY

> If an action is right (or wrong) for one agent in a certain circumstance, then it is right (or wrong) for any similar agent in similar circumstances.

This is one formulation of a non-substantive principle of ethical universalizability that has been given various labels: the Principle of Universalizability; the Generalization Principle; the Principle of Justice; the Principle of Impartiality; to mention the most common. In one variant formulation or another, this principle enjoys widespread acceptance among moral philosophers despite the fact that there is not widespread agreement about its meaning (and hence correct formulation); or about its justification (or how it should be "accounted for"); or about its implications, not only in connection with practical moral issues (i.e., concerning its practical application) but also in connection with other moral principles. These and associated issues are taken up in the papers by Singer, Rabinowicz, Nielsen, Olafson, Gorr, and Lycan and are thus grouped together in the anthology. In order to indicate how these authors approach these issues, it will be useful to provide brief sketches of their articles (though we have refrained from engaging in detailed exegesis and criticism).

The paper by Marcus Singer, 'Universalizability and the Generalization Principle,' touches on a number of the above-mentioned issues concerning this principle of ethical consistency. In particular, Singer considers its justification and various criticisms that have been leveled against it. The paper is divided into five sections. In Sections I and II, Singer considers certain important theses advanced by R. M. Hare concerning the principle's interpretation and justification. Hare prefers the label 'Principle of Universalizability' for the principle in question, and has argued on a number of occasions that, owing to important differences in meaning between 'universal' and 'general,' it is incorrect and misleading to speak of ethical generalization, or to use the label 'The Generalization Principle' (Singer's label) either for the principle in question or for associated ideas in ethics. In the first section of this paper Singer examines Hare's arguments for this claim and, finding them inconclusive, defends his use of 'generalization' for the idea and principle in question.

Hare has also advanced an account or explanation of the truth or validity of the principle of universalizability. According to Hare, Singular moral judgments (e.g., 'It is right (wrong, obligatory) for me to tell the truth in these circumstances') are universalizable in the sense that they commit the person making the judgment to a corresponding universal judgment ('It is (right, wrong, obligatory) for anyone who is relevantly similar to me in relevantly similar circumstances to tell the truth'). The entailment relation between singular and universal moral judgments holds because the principle of universalizability holds and the truth or validity of this principle is to be explained by reference to descriptive meaning rules that govern (at least in part) the meaning of such singular judgments. The central idea is this. On Hare's view, singular moral judgments in which such evaluative terms as 'right,' 'wrong', and 'obligatory' occur as predicates are analogous to ordinary singular descriptive judgments in which descriptive terms occur as predicates (e.g., "This is red"). The meanings of such predicates (both evaluative and descriptive) are governed by descriptive meaning rules-rules which, so Hare explains, "lay it down that we may apply an expression to objects which are similar to each other in certain respects." He adds, "It is a direct consequence of this that we cannot without consistency apply a descriptive term to one thing, and refuse to apply it to another similar thing (either exactly similar or similar in the relevant respects)" (Hare, 1963, p. 13). So, on Hare's view, the universalizability of moral judgments (or the principle of ethical universalizability) is to be explained or justified in a way analogous to the way the universalizability of singular descriptive judgments is to be explained, viz. by appealing to descriptive meaning rules.

In Section II, Singer criticizes this account of the principle of universalizability, charging that it leads not only to triviality but "leads [Hare] to move from his theoretical account of its basis—in descriptive meaning rules—to otherwise sensible judgments inconsistent with the original account."

In Section III, then, Singer turns to ills own account of the validity of this principle, reaffirming the view he offered in his *Generalization in Ethics*, according to which it is the inferential character of singular moral judgments—a feature that makes them analogous to "because" judgments in general and to causal and probability judgments in particular—that accounts for the validity of the generalization principle.

Turning to Sections IV and V, Singer (Section IV) considers the use Alan Gewirth has made of the generalization principle in the context of his own moral theory. Gewirth attempts to deduce what he claims

to be the supreme principle of morality—the so-called principle of generic consistency—in an argument one of whose premises is the generalization principle. Singer considers Gewirth's use of the principle and critically discusses some of the claims Gewirth has made about the generalization principle, in particular that it is neither a substantial nor a moral principle and that it "set no limits on the criteria of relevant similarity or the sufficient reasons for having the right to perform various actions" (Gewirth, 1978, p. 106). Singer argues in the first place that, when properly understood, this principle is both substantial and moral (as well as being logical) and furthermore, that "when properly understood [it] illuminates the concept of a genuine moral reason and what can count as one and what cannot, and shows that it is not open to anyone to invoke its form in justification of what is unjustified." Section V again takes up and considers further the criticism that the generalization principle sets no limits on what can count as a genuine moral reason.

The above formulation of a universalizability principle with which we began our discussion of the papers in this section makes reference to the notion of *similarity*. In fact, this notion which is absolutely crucial in this and related principles, admits of two different interpretations, one in terms of exact similarity, and the other in terms of *relevant* similarity. Each has special problems. The problem with the exact similarity formulation is that together with the Leibnizian thesis that indiscernability (exact similarity) entails identity, the principle turns out to be trivial. That is, together with Leibniz's principle, the ethical principle of universalizability thus interpreted reduces to: If an action is right (or wrong) for one agent, then it is right (or wrong) for *that* agent. On the other hand, the relevant similarities variant of the principle encounters the vexing (and, according to some, insurmountable) problem that the notion of relevance is unclear. Thus, we have what Rabinowicz calls the 'universalizability dilemma': embracing either of the only two alternative formulations seems unacceptable.

Rabinowicz's paper, 'The Universalizability Dilemma,' offers a way out of the dilemma. Making use of set-theoretic models and defining the principle of universalizability as a condition on a model, Rabinowicz employs, in addition to the concepts of exact similarity (copyhood) and relevant similarity, the concept of a *universal aspect* of a situation. Taking this newly introduced concept as primitive, Rabinowicz formulates a universalizability condition on a model that he claims is not trivialized by the Leibnizian principle and does not make use of the problematic concept of relevant similarity.

In 'Universalizability and the Commitment to Impartiality,' Kai Nielsen is concerned with the rationalist program of seeking ' the foundations of justice and sometimes the whole of morality in an *a priori* or formal principle *of Universalizability*." Nielsen denies that this program can be carried out and illustrates his claim by considering the attempts to go from the principle of universalizability (either directly or together with other considerations of logic and rationality) to an impartiality principle of the following sort: 'All human beings have an equal right to the fulfillment of their interests.'

Nielsen begins by making the same distinctions we have seen in Rabinowicz between the *exact* and the *relevant* similarity variants of the principle of universalizability. Working with the latter variant, Nielsen points out its two important limitations. First, it "does not tell us *what* is right or wrong, good or bad, or *what* ought or ought not be done. It says rather *if* one thing is right, good or ought to be done, then another thing relevantly similar to it is too." In other words, the principle is non-substantive. Second, Nielsen claims that the principle itself does not specify or set forth any criteria for determining what counts as being relevantly similar from the moral point of view. So, in order to go from the principle of universalizability to the principle of impartiality, one would have to judge first that some human being has a right to have his or her interests fulfilled and second that all human beings are (in general) relevantly similar. Since such additions do not follow directly from the principle in question and do not seem to be matters of logic, one can, without being inconsistent or irrational, affirm the principle of universalizability yet deny the principle of impartiality. (Nielsen reaches the same conclusion as a result of analysing Paul Taylor's claim that from the principle of universalizability one can derive the following principle of impartiality: 'Ifit is wrong for another to discriminate against him (the agent) on the ground of a difference he (the agent) does not acknowledge to be relevant, it must also be wrong for him (the agent) to discriminate against another on the ground of a difference the other does not accept as relevant.')

The upshot, at least if Nielsen is correct, is that doubt has been thrown on the rationalist strategy of using an apparently formal principle as a foundation on which to base substantive moral principles.

The papers by Olafson and Gorr are concerned with this general upshot of Nielsen's paper. Both papers consider the use of universalizability as a formal foundation (or at least as part of one), on which to base substantive moral principles.

As we have already noted, although Hare and Singer seem to disagree over the correct account or justification of the principle of ethical consistency, they do agree that this principle admits of a "logical" justification—one that makes essential reference to the logic of moral terms (see e.g., Singer, 1961, p. 34, and Hare, 1963, p. 30). We have also just noted Nielsen's argument that from such a formal, logically based principle, either alone or together with considerations of logic a.nd rationality, it is not possible to derive any substantive moral principles or judgments. In particular he argues that it is not possible to derive a principle of impartiality. The paper by Frederick Olafson, 'Reflections on a Passage in Mill's *Utilitarianism*,' offers an account of the principle in question that stands in sharp contrast to these views of Hare, Singer, and Nielsen.

Beginning with a suggestive passage from Mill's *Utilitarianism*, Olafson is concerned to defend a certain thesis concerning our allegedly rationally based commitment to a principle of impartiality (or what Olafson calls 'intersubjective reciprocity'). According to Olafson, "Mill's argument appears to postulate a direct connection between our social natures and a relationship to at least some other human beings in which we acknowledge an operative equivalence between their situation and our own." Olafson's elaboration of this argument in Mill goes as follows. We are essentially social creatures having a "social nature" involving, among other things, entering into agreements (both explicit and implicit) with other human beings. Such agreements presuppose "a set of conceptual instruments that function in a neutral manner as far as differences between persons and their points of view are concerned." This implies that in cases of conflicting interests, we must at least implicitly acknowledge the idea that, in general, no one person's interests occupy a special or privileged position. In other words, reflecting on our social natures reveals that we implicitly if not explicitly assent to the idea that, generally, the interests of all human beings are to be weighed equally. Therefore, a person who, as Olafson explains, takes or tries to take an egocentric point of view failing to weigh interests equally, "will not be able to persist in this course of conduct without feeling the strain of contradiction within the policies by which that conduct is governed." Thus, all conduct involving interpersonal transactions is governed by a set of presuppositions that make any social interaction possible, one of which is that of according equal weight to the interests of others. An egocentric line of conduct, however, denies this presupposition.

The relevance of all this for ethical universalizability, as Olafson explains, is that if we consider this principle in the "true context of ethical reflection ... that of an ethical community in which human beings stand to one another in determinate relationships that register the actual needs and circumstances and multiple forms of interdependency that in fact characterize our lives ... the requirement of universalizability is not simply an isolated logical thesis but widens into an immensely intricate and powerful dialectic of actions and persons." Thus, Olafson might agree with Nielsen that if we consider the principle of universalizability on its own as a logical thesis, then there is no contradiction involved in assenting to it yet denying a principle of impartiality. After all, as Nielsen points out, neither the principle itself nor considerations of logic and rationality entail the crucial judgment that all human beings are relevantly similar. However, analysis of "the pragmatic context" reveals that in general no one person's interests occupy a position of privilege vis-à-vis the interests of others, i.e., in general all human beings are relevantly similar when it comes to the satisfaction of their interests. This condition is a requirement that we are bound to accept on pain of contradiction, and, together with considerations of universalizability, it entails a principle of impartiality: In general, the interests of all other human beings are to be weighed equally. Thus, it is Olafson's contention that when interpreted in the proper context, and not simply as a logical thesis, the principle of universalizability does entail substantive conclusions.

In some of his recent writings, including *Moral Thinking: Its Levels, Point and Method*, R. M. Hare has attempted to accomplish what Nielsen argues cannot be done, viz., go from considerations of universalizability (together with non-moral principles of rationality) to a substantive moral principle. Hare's view is that "the requirement to universalize' our prescriptions generates utilitarianism" (Hare, 1981, p. 11). In 'Reason, Impartiality and Utilitarianism,' Michael Gorr investigates the connection in Hare's work between 'universal prescriptivism' (the name generally given to Hare's metaethical views) and utilitarianism.

Hare's argument for a kind of preference utilitarianism based on two formal properties of singular moral judgments—universalizability and prescriptivity—is roughly the following. The formal or logical properties of universalizability and prescriptivity impose certain formal rules for reasoning from the moral point of view. In particular, such features require what Gorr calls the "full reversibility test." That is, in judging from the point of view of morality one is to determine whether one can prescribe acting in accordance with a universal principle which,

as Gorr explains, involves determining "whether one would actually choose to perform that action if one knew that one would have to play, in a series of possible worlds otherwise identical to the actual world, the role of *each* person (including oneself) who would be affected." Furthermore, one is not "simply to imagine oneself with one's own interests, in the place of other persons," but, rather, to imagine "having in turn *their* interests and desires." This thoroughgoing impartiality requirement taken together with a Bayesian account of rationality entails the claim that a rational and fully informed universal prescriber would inevitably adopt and act on utilitarian principles. So, on Gorr's reconstruction, Hare's argument is this.

1. A rational person will always seek to maximize his expected utility. (Bayesian principle)
2. Utility = preference satisfaction.
3. Adopting the moral point of view is equivalent to deciding how to act as if one had (or were going to have) all the preferences of all those who would be affected by one's action. (Full Reversibility Test)
4. Therefore, a rational person who adopts the moral point of view will always seek to maximize the total expected utility of the group of all persons who would be affected by his action.

This move from universal prescriptivism to utilitarianism via the full reversibility test is, of course, hard to reconcile with Hare's well-known claim in *Freedom and Reason* that the thesis of universal prescriptivism is "normatively neutral" in the sense that it does not entail any moral theory or principle. However, as Gorr points out, there are traces of Hare's more recent views on the matter in *Freedom and Reason*, and, in the first part of his paper, Gorr suggests an explanation of why, in that earlier book, Hare held apparently inconsistent views on the question of the normative neutrality of universal prescriptivism. His suggestion is that there are two interpretations of the full reversibility test to be found in Hare's earlier writings (the "in *propria persona*" and "ideal observer" interpretations) which Hare did not clearly distinguish and which led him to opposing views on the normative significance of his metaethical views. What Gorr points out is that, in *Moral Thinking*, Hare has made clear that he now accepts the second interpretation of the full reversibility test and that this in effect ensures that principles chosen by rational and informed universal prescribers will be utility maximizing.

In the remaining sections of his paper, Gorr considers the plausibility of Hare's view that such a strong impartialism is constitutive of

morality (premise 3 of the above argument) and argues for the following claims: (1) The contention that a person might be motivated by a desire to be fully impartial is incapable of a non-circular explication; (2) Hare's claim that a strong impartiality condition is a formal requirement of morality is counterintuitive; and (3) Hare's metaethical theory does not entail such a strong impartiality requirement. If Gorr's objections go through, serious doubt has been cast on Hare's attempt to go from metaethical considerations concerning the logic of moral judgments (particularly the principle of universalizability) together with constraints on rationality to a substantive moral principle.

We have claimed that the principle under consideration is non-substantive in the sense that it does not alone (or even together with non-moral judgments) entail any substantive moral judgments. But to say this is not to say or even imply that this principle is useless or pointless or inapplicable in the context of practical moral issues.[4] Its use in such a context is illustrated by William Lycan in 'Abortion and the Civil Rights of Machines.' In fact it is one of Lycan's main contentions that an argument based on universalizability considerations is perhaps the best we may be able to do in making headway on the abortion issue. Lycan begins by claiming that because the fetus is a being of a "uniquely exceptional sort," we have no clear intuitions about the putative rights of such beings. Moreover, no further analysis of personhood or further fact finding is going to help us move closer to clarifying our intuitions on this matter. Thus, Lycan proposes that we explore this issue via a rather indirect route that involves reflecting on the status and rights of certain sorts of hypothetical machines. Lycan's proposed exploration proceeds roughly as follows.

Assuming that a materialist account of human beings is correct (that human beings have only physical attributes among their irreducible attributes) and in addition that it is technologically possible to (someday) construct a machine—a robot—that not only looks but behaves like an ordinary human being, Lycan claims that we could justifiably infer that such beings were conscious. Of course, given that consciousness of a certain complexity (involving certain cognitive and conative capacities) is sufficient for personhood, Lycan concludes that such sophisticated machines would be clear cases of persons. Lycan then goes on to consider the case of certain hypothetical machines—"proto-machines"—which do not yet possess those conceptual and conative capacities that are morally significant, but which are of a "highly individual basic design (which determines the personality it will have when it is finished)." We are to consider further the hypothetical case of

the proto-machine plugged into a "fully developed" mother machine and "ask whether it would be wrong for the technician to interrupt the mother-robot's activity at an early stage (with its concurrence if it is a person), unplug the proto-machine from the mother, take it off the workbench and dismantle it." Lycan reports no feelings of disapproval at the technician's actions and argues that since the proto-machine and the fetus appear to be relevantly similar, then by the principle of universalizability, we may conclude that at least early on in fetal development there is nothing morally wrong with aborting the fetus.[5]

KANTIAN UNIVERSALIZABILITY

The "universal law" formulation of Immanuel Kant's categorical imperative has already been cited in two different versions as Principles 3 and 4 near the beginning of this introduction. It was noted there that Kant made just about the most ambitious possible claim for this version of a universalizability principle, viz., that it is "the supreme principle of morality." Kant's view has been enormously controversial since it was first proposed, and issues of its interpretation and of its correctness have each generated an immense literature. The provocativeness of Kant's claim is added to by the fact that he believes that the principle is in no way a consequentialist one; he condemns all teleological, and hence all consequentialist views as "heteronomous" theories of morality. The papers in the present volume that center upon explicit discussions of Kantian universalizability are three. Onora O'Neill defends Kant's claims. George Nakhnikian finds them (and just about any other substantive universalizability principle) lacking. And Jonathan Harrison attempts a rapprochement between the historically hostile Kantian and utilitarian views; his paper could have been placed with equal appropriateness in either the section on Kantian or the one on consequentialist generalization.

The essay entitled 'Consistency in Action' by Onora O'Neill is a defense of the Kantian first formulation of the categorical imperative. She begins her essay (as does Nakhnikian) with an extended discussion of the Kantian concept of '"maxim." This is surely a key move for any would-be defender of Kant, for in the Kantian concept of the maxim as the formulation of the inner principle of action is where we find hope, if anywhere, of replying to objections that actions can be described in an indefinitely large variety of ways for the purpose of universalizing them. Any given individual action can be made to turn out to possess the quite different and incompatible moral qualities depending on the description one arbitrarily chooses to universalize. The idea is that the

maxim state a fundamental description under which the agent performs the action and this, at least when it can be formulated clearly and uniquely, provides the correct description of the action for the purposes of moral evaluation. O'Neill also makes the point that for Kant the primary moral quality of any action is an inner one of moral worth rather than an outward one of conformity with the requirements of external action. O'Neill, like Nakhnikian, emphasizes that there is a variety of ways in which maxims can be impossible or self-defeating prior to any consideration of their potential universalizability; but such difficulties must be cleared aside before a properly moral evaluation of the maxim can even begin. O'Neill states the intuitive idea behind the categorical imperative as being that" ... if we are to act as morally worthy beings we should not single ourselves out for special consideration or treatment" (p. 172). Actions which the universality test shows wrong are shown to result when we attempt to universalize acts or practices such as deception, coercion, and abrogation of autonomy. She states that much of the anti-Kantian literature here assumes that it is possible to state the proposed maxim of action in such a way that the action is described in quite a narrow fashion; but her earlier discussion of the nature of maxims has undercut this objection. The kinds of inconsistencies involved in Kant's "contradiction in the will" examples are also discussed. A point that is emphasized throughout O'Neill's essay is that Kant is not some sort of generalized utilitarian; the latter sort of view would be, according to Kant's terminology, a heteronomous moral theory. "The interest of an autonomous universality test is that it aims to ground an ethical theory on notions of consistency and rationality rather than upon consideration of desire and preference" (p. 182).

In a lengthy and thorough essay, George Nakhnikian considers very seriously the claim of the first formulation of the categorical imperative ("K1") to be what Kant calls it, "the supreme principle of morality," and finally rejects the claim. He looks with more favor upon the second formulation, which draws moral consequences from the fact that human beings are to be regarded as absolute ends who are not to be treated as mere means, and the third, "kingdom-of-ends" formulation which mentions as essential the peer status of moral agent-legislator-subjects. But our main interest in the present discussion is in Nakhnikian's thesis that "K1 does not work" (p.189).

As we consider this claim more closely we find that what Nakhnikian has to say has an interest quite far beyond issues of the exegesis of Kant. From its Kantian beginnings, his discussion becomes a much broader and more general discussion of the possible modes of univer-

salizability, in a sense attempting to cover the range of possible universalizability theses in the same thorough way that Narveson does, though as it turns out the possibilities that he discusses are mostly rather different from those discussed by Narveson. Thus for example, he discusses logically and physically impossible practical situations, and physically possible practical situations that cannot be universally practiced, and those that can be, but that in various ways we could not will to be universal practice.

The difficulties that Naknikian finds in Kantian and related versions of universalizability revolve around a set of counterexamples first stated on pp. 202-203, those of patenting an invention, visiting a sick friend in the hospital, publishing Icelandic jokes in a journal, and lying only in cases where no-one will find out.

Many discussions of universalizability criteria of morality have been carried on in terms of alleged counterexamples. Marcus Singer's *Generalization in Ethics* (1961) mentions a wealth of such proposed counterexamples and attempts to respond to them, and a fair amount of the literature discussing Kant on the application of the categorical imperative does the same. There is a large literature discussing Kant's famous four examples in the *Grundlegung zur Metaphysik der Sitten* (1785), the arguments for the wrongness of suicide, making a lying promise, and the obligation to develop one's talents and to render aid to another. These are not "counterexamples" to universalizability, at least certainly not so far as Kant's intentions were concerned! But Kant's application arguments have seemed to a number of writers either so inadequate or so obscure that for these people they have come to be regarded as counterexamples to Kant's claim to be able to apply the first formulation (the "universalizability" formulation) to moral situations to obtain unique correct answers to the question "Is an action of sort *S* morally correct or not?"

Nakhnikian sees his proposed counterexamples as not just showing the first formulation of the categorical imperative to be seriously defective. His conclusion is more general: it is that any such universalizability criterion must exhibit the same defects, must fall to the same counterexamples. Such scepticism about purely "formal" criteria of morality dates back at least to Hegel, who has some of his own proposed counterexamples.

On the other hand there also seems to be a perennial attractiveness about some concepts of universalizability, for they have their defenders today, as can be seen by examining the contributions to the present anthology by Narveson, O'Neill, and Singer, among others.

The counterexamples that Nakhnikian proposes call for further detailed discussion, something that is beyond the scope of this introduction. One large class of counterexamples to Kantian universalizability deal with the alleged arbitrariness of the deSCription of actions for the purpose of moral evaluation. Singer's discussion of the concepts of "invertibility" and "reiterability" in *Generalization in Ethics* (1961) is one important attempt to deal with this problem. Potter (1973) also discusses this issue in its Kantian context. There are no doubt also other classes of counterexamples that result from other kinds of problems with universalizability, such as questions about the relevance or irrelevance of the wants, wishes, or likely actions of others. (Is it right to perform an action that is such that if many performed it, we could not accept the consequences, just in case very few others are likely to want to perform such an action? Is it obligatory to render aid to another when others are failing to do their part in rendering aid, and this makes it either likely or certain that my own attempt to render aid will be unsuccessful?) Another kind of issue arises when we are considering cases of actions based on false beliefs, a kind of case that Kant never mentions, and still different cases arise from puzzles concerning competition, in which, necessarily, not everyone can win (O'Neill discusses these in her paper).

CONSEQUENTIALIST UNIVERSALIZABILITY

As has already been remarked, the paper by Jonathan Harrison, though placed by us in the division on consequentialism could equally well have been placed in the section on Kantian universalizability. Harrison's main aim is to show that there are large areas of agreement between Kantians and utilitarians, and in particular that many of Kant's criticisms of teleological ethics do not apply as sound criticisms to utilitarianism.

The version of utilitarianism that Harrison defends is a version he calls "cumulative effect utilitarianism," to distinguish it from rule utilitarianism, and a kind of "ideal" utilitarianism which is prepared to allow that Kantian moral virtue might be something good in itself, an admission that would of course be incompatible with the hedonism of classical utilitarianism. Here are some of the Kantian objections that in Harrison's view do not have proper application to utilitarianism: (1) Kant thought that teleological theories all appeal to the agent's own inclination, but utilitarianism, which demands that we seek for the good of all, does not. We are not likely to have an inclination to bring about the good of all, and it is also clear that our having the utilitarian obligation does not pre-

suppose our having the inclination. (2) It is quite possible and reasonable to develop a version of utilitarianism which does not have the rightness or wrongness of acts depend on actual consequences, but instead has such moral qualities of acts depend on intended consequences and the qualities of character which tend to produce certain kinds of facts. Such a version of utilitarianism is closer to Kant's view. (3) Kant thought that duty, as a modal notion, implied a kind of necessity, and he thought that teleological ethics was incompatible with morality properly conceived. Harrison in response tries to sort out the kinds of necessity which moral judgments do and do not have, mentioning some of the possible confusions that may arise concerning such issues. He concludes that Kant may have been guilty of certain confusions, and that in any case Harrison's version of utilitarianism is not incompatible with the proper kinds of necessity attaching to moral judgments.

Meanwhile, Harrison does have some criticisms of Kant: (1) He thinks that Kant exaggerated the goodness of good will. Even if it is allowed as a good in itself, it is not the only good in itself, nor is its good infinitely greater than other kinds of goodness. (2) On the matter of greatest interest in this anthology Harrison says that Kant's universal law formulation of the categorical imperative marks "the point at which utilitarianism has to be modified in order to fulfill Kant's requirements." The modification arises in cases such as promise-keeping, where the utilitarian would recommend failing to keep promises in certain individual cases; if such recommendations were commonly followed by bad consequences—the cumulative consequences of a large number of promise-breakings—would follow. So the individual is well advised not to follow the utilitarian's advice, not because his maxim would be self defeating if universalized (this fact has no relevance to the moral quality of the act in Harrison's view), but because there would be bad consequences. Harrison recognizes that this point does not quite mean that he is accepting Kants first formulation, but he thinks it comes rather close.

The papers by Cork and Gillespie concern Marcus Singer's generalization argument:

> If everyone were to do that, the consequences would be undesirable, therefore no one ought to do that.

The principle corresponding to this argument,

> If the consequences of everyone's doing that would be undesirable, then no one ought to do that

functions in Singer's theory as the supreme moral principle and thus

"serves as a test or criterion of the morality of conduct, and provides the basis for moral rules" (Singer, 1961, p. 9).

Now Singer offers an argument or deduction of this principle (GA, hereafter) of the following sort.

(1) If the consequences of *A*'s doing *x* would be undesirable, then *A* ought not to do *x*. (The principle of consequences, C.)
(2) If the consequences of everyone's doing *x* would be undesir able, then not everyone ought to do *x*. (The generalization from the principle of consequences, GC.)
(3) If not everyone ought to do *x*, then no one ought to do *x*. (The generalization principle, GP.)
(4) Therefore, if the consequences of everyone's doing *x* would be undesirable, then no one ought to do *x*. (GA.)

Premise (2) is said to be a generalization from (1) –the latter taken to be necessary—and (4) is supposed to follow from (2) and (3).

This argument has been the subject of a good deal of criticism, perhaps the most common being the so-called 'collective-distributive' criticism which can be explained as follows. The term 'everyone' can be construed *collectively* as referring to the whole of a class of (relevantly similar) agents or *distributively* as referring to each and every particular member of that class. Accordingly, GC admits of two interpretations depending on how 'everyone' is understood in the consequent.

> If the consequences of everyone's doing *x* would be undesirable, then it ought to be that someone not do *x*. (The collective interpretation.)
> If the consequences of everyone's doing *x* would be undesirable, then there is someone who ought not to do *x*. (The distributive interpretation.)

The problem is this. On the collective interpretation, the argument is invalid since as Singer makes clear 'everyone' in the antecedent of GP (Premise 4) is to be read distributively. However, on the distributive interpretation, GC is susceptible to counterexamples of the following sort. Although if no one stands guard duty on a particular occasion, the consequences would be undesirable, it does not follow that there js someone about whom we could say that *he* ought to stand guard duty and could be punished for not doing so (Nakhnikian, 1964, p. 446)

Charles M. Cork, in his 'The Deontic Structure of the Generalization Argument', develops a formalized version of Singer's argument that attempts to avoid this collective-distributive problem. His new version involves an analysis of the consequent of GC ('it ought to be the case that someone not do *x*') that allows a transition to a distributive claim about

the members of the relevant set of agents but "not to those flesh and blood relevantly similar agents ... to whom we can point, but to logical constructions of agents (blanks, as it were), into which the former can be filled ...." In other words, instead of distributing the collective obligation to avoid undesirable consequences to identifiable persons (which then raises the embarrassing question of who is forbidden to perform the action in question), on Cork's analysis, such obligation is distributed to "*arbitrary members* of a set, .. defined here as those about whom we make no assumptions other than their membership in the set."

If successful, Cork's analysis of this argument not only solves the collective-distributive problem but helps make clear how the various alleged counterexamples to the generalization argument can be handled by Singer's theory.

Antecedent to the publication of David Lyons' *Forms and Limits of Utilitarianism*, it was generally thought that so-called *general* utilitarianism (a label used to characterize Singer's generalization argument) was immune to certain apparently devastating criticisms besetting analogous forms of *simple* or act utilitarianism. The claim was that a principle like Singer's is able to account for our obligations concerning generally accepted moral rules that are troublesome on the act utilitarian view. This alleged difference between analogous forms of these two sorts of theory was challenged by Lyons who argued persuasively that despite appearances, analogous forms of simple and general utilitarianism are really extensionally equivalent—they yield the same substantive judgments when applied to the same cases. In other words, according to Lyons once we bring into consideration all of the relevant facts and utilities bearing on a particular case, "it matters not whether we ask 'What would happen if everyone did the same?' instead of 'What would happen if this act were performed?,'" (Lyons, 1965, p. 119), the results of these two tests will be the same.

Our final article, Norman Gillespie's 'Moral Reasons and the Generalization Test in Ethics', concerns Lyons' extensional equivalence thesis. In particular, Gillespie considers Lyons' explanation of the relevance of the behavior of others in describing actions for purposes of applying the generalization test which is the key to the reductive thesis. Gillespie distinguishes two readings of the generalization test—the *de dicto* and the *de re*—and argues that it is Lyons' *de dicto* reading of the test that leads him to mistaken conclusions about the relevance of the behavior of others in applying the test. According to Gillespie, once the test is interpreted *de re*, which "is the appropriate one for capturing its moral force," we can see not only that the extensional equivalence

thesis fails but that the test is not strictly utilitarian. In fact, as Gillespie explains, "The *moral* point of the generalization test in ethics is *not* that your act will, either incidentally or as part of a general practice, produce undesirable consequences, but that it is unfair."

CONCLUSION

The aim of the present volume is to show that recent philosophical thought on universalizability is multifaceted and alive, and is making advances. This has been done by presenting a wide variety of work by authors who are among the best of those currently working on these issues. This introduction has aimed to present the main points of our authors in summary form, exhibiting some of their relationships to each other.

Now that we have briefly discussed the papers included here it can be seen that they each have an important role to play in further discussions of this general topic. The pieces by Singer, Gillespie, and Cork discuss and reassess one of the major works on this aspect of moral philosophy, Singer's own *Generalization in Ethics*. O'Neill, Nakhnikian, and Harrison take views each opposed to the other on the interpretation of Immanuel Kant's view on universalizability. Kant is surely the single most important historical writer on this topic in ethics. Michael Gorr takes up a recent important new work by another major contemporary figure in the literature of universalizability, the man who created the term, R. M. Hare. Nielsen considers (and rejects) connections with the seemingly related concept of impartiality. Lycan considers a problem arising in the application of certain universalizability principles, viz., that some kinds of moral issues seem to be unique, and thus, frustratingly, to allow of no clearcut moral analogies with other issues that could be used to alleviate our moral perplexity. Rabinowicz discusses a closely related issue that is also discussed in various ways by O'Neill and others, the problem of how to interpret the universalizability principle so as to avoid the Scylla of the banal useless truism, and the Charybdis of an unuseable notion of "ethical relevance." The issue is that of the relativity of descriptions of action that are to be evaluated morally, the multiplicity of available descriptions, and the seeming arbitrariness of choosing one description over another.

The essay by Frederick Olafson, which chooses as its text an interesting and neglected passage from an important classic of moral philosophy, J. S. Mill's *Utilitarianism*, illustrates the ways in which utilitarian and deontological themes often intermingle in discussions of universalizability. The same point is illustrated in different ways in the work here by Harrison, Gillespie, and Narveson. "Unfairness" considerations are

surely best understood as nonconsequentialist if not Kantian in their appeal. The fact that (if Gillespie is correct) Singer's theory, which has been taken by most readers to be consequentialist, contains, when properly understood, an important appeal to the deontological characteristic of "unfairness" illustrates the complex interplay of consequentialist and Kantian moral characteristics which we have also seen in other papers.

In fact it is often difficult to discern when a conception of universalizability is purely consequentialist as opposed to when it contains appeals only to deontological elements. Likewise it is sometimes difficult to discern whether a universalizability principle is merely "logical" and hence nonsubstantive, or whether it may in fact have some important element of extra-logical substance.

The distinctions on which we based the organization of this volume of essays, though important ones, are often overlooked or overridden by actual philosophical practice. This present introduction is not intended as a brief for philosophical purity with respect to such distinctions. The common presence of "impurity" surely reflects something important about the moral phenomena that are under discussion.

There is no common theme or new consensus that emerges from this collection. Any attempt to achieve consensus or to unify under a common theme is surely premature. In no case, we think, does an essay in this collection resolve and thereby close off further philosophical discussion. The usefulness of the essays is much more likely to lie in a different direction—in the direction of providing advances upon previous current discussion and thereby not closing off but rather opening up these issues and providing a stimulus to further discussion.

NOTES

[1] Aristotle, *Nicomachean Ethics* (Ostwald translation (Indianapolis: Bobbs-Merrill, 1962, p. 118), Book V, Chapter 3, 1131a13-23).

[2] An even more radical thesis would be that of giving the treatment that will yield equal results (compensating equality). For a discussion of this sort of equality see Onora O'Neill s essay, 'Opportunities, Equalities, and Education' (*Theory and Decision*, 7 275-295, October 1976).

[3] Without involving ourselves in the task of defining what counts as a consequentialist moral theory or principle, suffice it to say that we are using the term in a general way intended to cover utilitarians of all varieties as well as, e.g., Singer's theory which, as he correctly insists, is not, strictly speaking, utilitarian (see Singer, 1977).

[4] On this point see Sid B. Thomas (1968).

[5] As Lycan admits, the analogy here needs some adjusting which he attempts to supply and, moreover, this particular argument is only persuasive (if at all) when considering the proto-machine at an early stage of construction development.