NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

FOUNDATIONS OF UTILITARIANISM

by

Robert W. Bright

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
June, 1991

ISBN 0-315-71490-5

for Dorothy and Carolyn

TABLE OF CONTENTS

ABSTRACT

The problem of interpersonal utility comparisons has remained a thorn in the side of utilitarian theorizing for over a century. The present work argues that the problem rests on a metaphysical assumption: namely, that preferences bear their intensities as monadic properties, so that it makes sense to attempt to directly compare the strength of some of an individual's preferences with those of someone else, without regard for the strengths of their remaining preferences. The failure of behavioural and physiological proposals for performing such "direct" comparisons is carefully examined, and further considerations are adduced to support the view that preference intensities are relational rather than monadic properties. A proportionate interpretation of individual welfare is then sketched which forms the basis for "indirect" interpersonal comparisons, on which preference strengths are compared via their proportionate contribution to each person's total possible welfare. The final chapters argue that the usual distributive objections to utilitarianism cannot be sustained against a version based on indirect comparisons; and that recent complaints centered around the notion that utilitarianism is too demanding as a moral theory depend on misconstruing it as a theory of obligation rather than as a theory of the moral worth of actions.

CHAPTER 1: INTRODUCTION

Utilitarianism requires the comparison of different individuals' welfares, and such comparisons are far from unproblematic. The problem of interpersonal utility comparisons has plagued political philosophers and welfare economists for over a century. Despite the considerable attention which the subject has received, particularly in the years since Lord Robbins launched his famous attack on the possibility of comparing the satisfactions of different individuals, nothing like a settled opinion has emerged concerning whether or to what extent objective welfare comparisons are possible.

A quick inventory of the literature on interpersonal comparisons reveals that the problem has received rather more attention from economists and other social scientists than from philosophers. With a few exceptions, the latter have by and large contented themselves with hinting at some of the difficulties involved in performing welfare comparisons, while leaving fuller discussions to their colleagues in other disciplines. Rawls' treatment of interpersonal comparisons is fairly typical in this regard; in discussing the advantages of his two principles of justice over the principle of utility, he writes: "I do not wish to stress these much discussed technical problems, since the more important objections to utilitarianism are at another

level."[1]

From the point of view of the present work, the attitude towards utilitarianism and interpersonal comparisons encapsulated in this brief quotation from Rawls is misguided in two respects. In the first place, so far as utilitarianism is concerned, the problem of interpersonal comparisons must be regarded not so much as a technical problem than as a foundational one. Utilitarianism depends critically on the ability to measure and compare welfare across individuals; without some account of utility comparisons, the principle of utility is virtually without content. Moreover, as we shall see, both the form of the utilitarian calculus and its distributive consequences are inextricably bound up with the kind of comparability that is licensed by a satisfactory resolution to the problem of interpersonal comparisons. Rawls' words suggest that interpersonal comparisons can be viewed somewhat in the manner of a technical adjunct to utilitarian theory proper, and that the theory can be evaluated independently of the specific character of this adjunct. This view of the relation between utilitarianism and welfare comparisons is

---

[1] A Theory of Justice (Cambridge, Mass.: Harvard University Press, 1971), p. 321. For some doubts about whether Rawls' difference principle really does possess the advantages over the principle of utility which he goes on to cite, see Kenneth J. Arrow, "Some Ordinalist-Utilitarian Notes on Rawls's Theory of Justice", Journal of Philosophy 70 (1973), pp. 245-63; reprinted in Collected Papers Volume 1: Social Choice and Justice (Basil Blackwell, 1984), pp. 96-114.

very likely the prevailing one among philosophers; in the absence of a solution to the problem of interpersonal comparisons, it is of course indispensable for getting on with the business of serious ethical theorizing. Nevertheless, the view in question is fundamentally mistaken. Welfare comparisons comprise the very core of utilitarianism, not just in the sense that the theory itself comes to naught in the absence of interpersonal comparisons, but also in the sense that a great many objections to it (including those which Rawls regards as "the more important objections") stand or fall largely on the form which interpersonal comparisons may take.

Secondly, there are real grounds for doubting that the problem of interpersonal comparisons may be fruitfully regarded as an essentially "technical" problem, even when considered apart from utilitarianism. Rawls presumably refers to the difficulties which he goes on to cite concerning the measurement and comparison of welfare as technical problems because recent discussion of them has taken place primarily within contexts where the formal procedures of utility theory and the theory of social choice are the participants' stock in trade. Notwithstanding recent trends, however, early discussions of interpersonal comparability did not revolve around technical <u>desiderata</u>. Nor, in all fairness, do formal techniques appear to have rendered the problem any more

tractable.

This continuing recalcitrance in light of formal developments highlights the need for a reconsideration of the philosophic grounds and implications of the problem of interpersonal comparisons. In the next chapter and the one following I shall argue that the problem does not in any important sense rest on formal considerations. Nor, in spite of the fact that the technical apparatus of utility and social choice theory provides a convenient and sometimes revealing framework within which to formulate certain aspects of the problem, does a solution to it depend upon invoking this apparatus. Rather, the problem is generated by certain deep metaphysical presuppositions concerning the nature of individual welfare, and only a careful analysis of these presuppositions can point the way towards a satisfactory resolution. So in this sense too, the problem of interpersonal comparisons should be regarded as a foundational problem rather than a technical one.

More specifically, what I will be arguing in the next few chapters is that the problem of interpersonal utility comparisons is rooted in a mistaken conception of preference intensity. Without explicitly addressing the issue, most parties to the debate over interpersonal comparisons have assumed that preference intensities are monadic properties of preferences; i.e., that a given preference possesses its strength independently of the strengths of

the other preferences with which it resides in an individual's preference set. The focus of the debate has thus been on the question of whether there is some empirical test by means of which we can directly compare the intensity of some preference of A's with the intensity of a preference of B's, without regard for the strength of their remaining preferences. I refer to attempted comparisons of this sort as direct interpersonal comparisons.[2]

Following some preliminary remarks concerning utilitarianism and the problem of interpersonal comparisons at the beginning of Chapter 2, we examine two well-known proposals for arriving at direct comparisons of individuals' welfares. Our conclusion will be a familiar one: that direct utility comparisons are not, objectively speaking, on the cards. Our motivation for discussing these proposals is somewhat novel, however; we will be less concerned with establishing that direct interpersonal comparisons fail, than with examining how they fail. Careful consideration of this latter point will provide us

---

[2]This choice of terms is not an entirely happy one. Of the two proposals for performing direct interpersonal comparisons which we will consider in Chapter 2, one defends comparisons which are in a perfectly good though unrelated sense more "direct" than the other. I.e., the physiological comparisons defended by Richard Brandt are in a sense more direct than John Harsanyi's behavioural comparisons, since the physiological structures which form the basis for preferences and preference intensities are also the structures which in part determine behaviour. In our sense of the term, however, Brandt's and Harsanyi's proposals are equally concerned with the possibility of performing "direct" comparisons.

with some important hints concerning the nature of preference intensity.

Chapter 3 is devoted to exploring the hypothesis that direct interpersonal comparisons fail because preference intensities are not monadic properties at all, but are rather irreducibly relational properties of the preferences which possess them. Drawing on an analogy with multi-tasking computers, I argue that preference intensities are best understood as relations between the preferences within an individual's preference set which serve to regulate her behaviour given the need to allocate scarce but variable amounts of resources in satisfying those preferences. I then proceed to defend the implications of this relational view of preference intensity for what I take to be an adequate conception of individual welfare. And finally, in light of our discussion of the relational nature of preference intensity and individual welfare, I offer a solution to the problem of interpersonal comparisons which rests on the possibility of performing indirect interpersonal comparisons, comparisons on which statements of the form "A's preference for x over y is stronger than B's preference for z over w" are understood as being elliptical for "A's preference for x over y means more to her (it occupies a greater space on her own personal scale of value) than B's preference for z over w means to him." The chapter is rounded out with a discussion of the role that

indirect comparisons play in grounding the so-called "ordinal" comparisons which have recently received attention in the literature on social choice.

In Chapter 4 we direct our attention to the consequences for utilitarian theorizing of adopting the solution to the problem of interpersonal comparisons developed in the preceding chapter. As suggested above, indirect comparisons will do far more than simply rescue utilitarianism from the charge that it rests on questionable appeals to utility comparisons. Indirect comparisons also shape the form and content of utilitarian prescriptions in a distinctive and intuitively appealing fashion, and thereby insulate the theory from a range of otherwise powerful objections. At the purely formal level, Rawls' complaint that utilitarianism "does not take seriously the distinction between persons" cannot be sustained against a utilitarianism based on indirect utility comparisons. Less formally, indirect comparisons place severe distributional constraints on utilitarian directives. It is a consequence of the account of welfare and interpersonal comparisons defended in Chapter 3 that individuals cannot diverge in their overall capacities for satisfaction. It follows that "utility monster" types of objections are not formulable against a utilitarianism based on indirect utility comparisons; and more generally, that utilitarian prescriptions are far more egalitarian than has commonly been supposed.

What holds in theory here carries over nicely to application: In the final section of Chapter 4 I argue that the ordinary concept of needs as it plays a role in guiding ethical practice is plausibly understood as arising directly out of our practical inability to compare welfares in detail; and hence that utilitarianism is capable of furnishing badly-needed underpinnings for real-world social policies directed towards insuring that individuals' needs are met.

Indirect interpersonal comparisons thus provide the foundation for an account of utilitarianism which is more detailed, more coherent, and eminently more defensible than the often confused interpretations which have borne the brunt of recent criticism. My final order of business, in Chapter 5, will be to sketch a utilitarian theory of obligation with which to supplement the general theory developed in the preceding chapters. Bernard Williams among others has criticized utilitarianism for being too "simple-minded" a theory to be able to deal satisfactorily with the complexities of moral and political life. On Williams' view, utilitarianism's lack of conceptual resources blinds it to concerns, such as equality and personal integrity, which manifestly cannot be ignored in real moral and political decision-making.[3] I shall argue

---

[3]See "A Critique of Utilitarianism" in J.J.C. Smart and Bernard Williams, Utilitarianism: For and Against (Cambridge University Press, 1973), esp. pp. 149-50.

that such objections are misplaced, in part for reasons already addressed by the end of Chapter 4, but also in large part because they badly misconstrue the structure of consequentialist thinking in general and utilitarian thinking in particular. The objections are directed not so much towards utilitarianism's simple-mindedness as towards a simple-minded interpretation of the theory, one which fails to maintain an adequate separation between the utilitarian's central account of the moral value of actions and her account of obligation. Once this separation is secured, the way is open to providing a non-maximizing account of obligation which makes room for supererogatory actions, appropriately locates concerns such as personal integrity within the utilitarian framework, and sheds a good deal of light on what it means from the subjective point of view to lead a characteristically moral life. Thus when the overall structure and foundations of the theory are clearly perceived, utilitarianism provides us not only with a powerful account of distributive justice in the abstract, but is as capable of attending to the complexities of moral and political decision-making as any theory can and ought to be.

CHAPTER 2: THE PROBLEM OF DIRECT INTERPERSONAL COMPARISONS

1. Utilitarianism and Welfare

In its most general form, utilitarianism is the ethical and political doctrine which asserts that the moral worth of actions, intentions, dispositions, policies, institutions, etc. is determined solely by the extent to which they promote aggregate utility or welfare.[1]

In characterizing utilitarianism in this fashion, I mean to be emphasizing the central importance which it assigns to moral rankings of actions etc. in a way which runs against the grain of received practice. Proponents as well as critics of the doctrine have often conceived of utilitarianism in the first instance as a theory of obligation: as the theory, namely, which obliges one to maximize aggregate welfare. Thus J.J.C. Smart tells us

---

[1]More precisely, since choice and evaluation typically take place in circumstances where knowledge is less than perfect, utilitarianism evaluates actions, intentions, etc. solely by the extent to which they promote aggregate expected welfare, with individuals' expected welfare defined in the usual way, as the product of their welfare consequent on possible outcomes times the probability of those outcomes. (Note that Bentham explicitly provided for expected welfare in his formulation of utilitarianism in citing "certainty or uncertainty" as one of the four primary "dimensions of value" of pleasures and pains; see An Introduction to the Principles of Morals and Legislation, Oxford: Clarendon Press, 1823, p. 29.)

I shall be using the terms 'utility' and 'welfare' to refer indifferently to whatever it is that the utilitarian holds to be the ground of moral worth. On reasons for thinking that some conceptions of welfare are better suited to utilitarian purposes than others, see below.

that act utilitarianism is "the view that the rightness or wrongness of an action is to be judged by the consequences, good or bad, of the action itself", and this is quickly abbreviated to the maxim "maximize probable benefit".[2] In a more general vein, Rawls defines teleological theories (of which utilitarianism is supposed to be a paradigm) as theories on which "the good is defined independently from the right, and then the right is defined as that which maximizes the good."[3] After some grappling, Williams finally distinguishes consequentialist positions from non-consequentialist ones in virtue of the supposed fact that on the former but not the latter, an action's being right entails that an optimific state of affairs has resulted.[4] And Sen, with his usual clarity, characterizes "utilitarian moral structures" as combining the central element of "outcome utilitarianism" (which holds that one state of affairs is at least as good as another if and only if the sum of individual utilities is at least as large in the former as the latter) with some "consequentialist" principle asserting that an action, rule, or whatever is right if and only if it would result in a state of affairs at

---

[2] "An Outline of a System of Utilitarian Ethics" in Smart & Williams, op. cit., pp. 9 and 12.

[3] A Theory of Justice, p. 24. The definition is borrowed in essentials from William K. Frankena, Ethics (Englewood Cliffs, N.J.: Prentice Hall, Inc., 1963), p. 13.

[4] "A Critique of Utilitarianism", pp. 82-93.

least as good as any of the available alternatives.[5]

Despite the neatness of the formulas, these ways of characterizing utilitarianism are in my view seriously misleading. In alleging that the utilitarian's immediate concern is with providing a theory of right and wrong actions (policies, rules, institutions, whatever), or at any rate that her theory of obligation follows immediately from her specification of the good (and hence that for practical purposes we may simply define the general theory in terms of an obligation to maximize the good), they inflate the importance of deontological concepts in a manner which threatens to undermine utilitarianism's status as a distinctively consequentialist theory. I shall return to this issue in Chapter 5, arguing that for the utilitarian what one is obliged to do should be understood as distinct from and wholly subsidiary to the question of what it would be morally better for one to do. For now, we will simply note that Bentham's characterization of the principle of utility at the outset of the Principles makes no mention of obligations; it is:

> that principle which approves or disapproves of every action whatsoever, according to the tendency which it appears to have to augment or diminish the happiness of the party whose interest is in question: or, what is the same thing in other words, to promote or to oppose

---

[5]Amartya Sen, "Utilitarianism and Welfarism", Journal of Philosophy Vol. LXXVI, No. 9 (1979), pp. 463-89.

that happiness.[6]

In a nutshell, then: For utilitarians, the more welfare resulting from an action, policy, etc., the better.

The principle of utility is not completely contentless in the absence of interpersonal comparisons of welfare. Within a given set of contemplated actions or policies there may be some alternatives which are Pareto-superior to others -- i.e., some actions or policies which would result in at least one person being better off, and no-one being worse off, as compared with some of the other alte- t-ives.[7] If one alternative is Pareto-superior to . other then it is unequivocally superior from the standpoint of

---

[6]Ibid., p. 2. Bentham attends briefly to the deontological concepts of 'right', 'wrong', and 'ought' eight paragraphs hence, but only after making it clear that the principle of utility is to be understood in the first instance as a principle by which to rank the moral worth of actions, as e.g. in the immediately preceding paragraph:

> A man may be said to be a partisan of the principle of utility, when the approbation or disapprobation he annexes to any action, or to any measure, is determined by and proportioned to the tendency which he conceives it to have to augment or to diminish the happiness of the community.... (pp. 3-4, my emphasis)

[7]Strictly speaking this is a characterization of "strong" Pareto-superiority. The weak version of the Pareto principle (so-called because it is implied by the strong version) states that one action is superior to another if the former would result in everyone being better off as compared with the latter. (See Amartya K. Sen, Collective Choice and Social Welfare, San Francisco: Holden-Day, Inc., 1970, p. 24.) Thus, in one of those terminological twists common to formal theorizing, if at least one person would be better off as a result of adopting policy x rather than policy y, and no-one would be worse off, then x is strongly superior to y; whereas if everyone is better off under x rather than y, x is weakly superior.

promoting aggregate welfare, and hence provides the utilitarian with a limited basis for moral ranking, irrespective of how (if at all) welfare may be compared across individuals. The basis is likely to prove rather slim, however. For any action or policy which is Pareto-superior to another, there will almost inevitably be still other alternatives to which the first is not Pareto-superior; there are few if any real-world situations in which, for a given course of action, there is not at least one alternative on which someone could be made better off at the expense of making others (perhaps everyone else) worse off. Thus in all but the most exceptional of circumstances we will not be able to specify which among a set of available actions or policies is the utilitarian-preferred one, unless we can discover some principle by means of which to weigh the welfare gains of some individuals against the welfare losses of others.

Before turning to some of the issues surrounding interpersonal comparability, we should briefly attend to a prior concern regarding the notion of welfare itself. If utilitarians are united in citing welfare as the sole determinant of moral worth, they have not quite been univocal on the question of how welfare is to be understood. On the received interpretation, early utilitarians such as Bentham and J.S. Mill were largely "hedonists"; i.e., they conceived of welfare in terms of the <u>felt</u>

quantities of pleasure and pain which an individual experiences during some interval. Recent theorists on the other hand have mostly favored the view that welfare is the satisfaction of preference or desire, with 'preference' and 'desire' construed in the widest possible terms to encompass whatever it is that individuals may care about.[8]

James Griffin has recently undertaken a searching examination of some of the issues involved in the choice

---

[8]A brief note on the relation between preference and desire: At first glance one might suppose that grouping the satisfaction of preferences and the satisfaction of desires together under the same banner results in the conflation of two quite distinct concepts. Desires appear to be dyadic relations between a desirer and an object or state of affairs which is desired: 'Individual A desires x'. (Philosophers who categorize desires as a variety of propositional attitude take statements of this form to be elliptical for 'A desires that she possess x', or 'that x be the case'.) Preferences on the other hand are clearly three-part relations between an individual and two objects or states of affairs; one cannot strictly speaking have a preference for x simpliciter, one can only prefer x to something else y.

This apparent difference should not be taken to mark a fundamental distinction between preferences and desires, however. The key to understanding the relation between the two lies in noticing that people do not, at least in the ordinary way of speaking, desire things which they already possess. (True, I may sensibly ask someone if she wants, say, the apple which is already in her possession; but this is shorthand for asking her if she wants to retain the apple rather than giving it to me, and keeping the apple is not something which she already possesses.) Thus in ascribing desires to people we make implicit reference to the status quo: If someone desires x, this is just to say that she prefers x to the status quo (more precisely, that she prefers having x to not having x, other features of the status quo remaining constant so far as possible). A person's desires, then, are simply a subset of her preferences as a whole (though particularly important ones from the standpoint of explaining her behaviour, since it is precisely individuals' preferences vis-a-vis the status quo which motivate them to act).

between preference-based and pleasure-based accounts of welfare, and I am in substantial agreement with much of what he has to say concerning the inadequacy of the latter.[9] I shall limit myself here to emphasizing a point that Griffin does not: namely, that preference-based accounts of welfare are much better suited to expressing certain of the utilitarian's most fundamental convictions than hedonistic accounts are.

Utilitarians have always been deeply motivated by a desire to divest ethics of intuitionistic precepts and ground moral value firmly in the empirical world. If things have moral value at all, according to the utilitarian, it is not because God or any other authority says

---

[9]See Well-Being: Its Meaning, Measurement, and Moral Importance, (Oxford: Clarendon Press, 1986), Part One. I cannot, however, endorse Griffin's own desire-based account of welfare, for two reasons: (1) He attempts to narrow the scope of his account so that the satisfaction of some of an individual's desires does not count towards her welfare (roughly speaking, those desires which are remote in the sense of not occupying a central place in her overall life plan, though the criteria for what gets ruled out and what doesn't are less than clear). (2) His account is tinged with a kind of objectivism which rests on what appear to me to be extravagant claims concerning the relation between "understanding" and desire; for a diagnosis of Griffin's error in this regard, see R.M. Hare's response (Hare and Critics: Essays on Moral Thinking, Oxford: Clarendon Press, 1988, pp. 234-41) to Griffin's "Well-being and its Interpersonal Comparability" (ibid., pp. 73-88). Part of the difference between us may be attributed to the fact that Griffin is attempting to systematize the ordinary concept of welfare or well-being (whether there is such a concept suitable for unified treatment, given recent connotations of the term, is a question I pass over), whereas my sole concern is to arrive at a conception of welfare which yields an overall best fit with utilitarian convictions and purposes.

that they do. Nor can the existence of moral value rest on the supposed existence of faculties such as Moral Sense or Understanding or on the Law of Nature (which is Reason); far from providing a ground of value, these metaphysically and epistemologically suspect categories only mask a refusal to provide such grounds.[10] Value gets into the world if it gets there at all, on the utilitarian's view, in virtue of the fact that there are people in the world who actually value things. Thus whatever distinctively moral value we find in the world must ultimately be traceable to whatever it is that individuals value.

Preference-based accounts of welfare provide a more robust expression of this conviction than pleasure-based accounts do for the simple reason that people may place a significant degree of value on things other than pleasure. Niceties aside, the issue between the two types of account turns on whether (i) preference-based accounts really do diverge from hedonistic accounts in important respects; and if so, whether (ii) the divergence is such as to legislate in favor of one sort of account over the other.

There is little doubt that pleasure and desire do part company in ordinary speech.[11] Appeals to ordinary language

---

[10]Cf. Bentham, ibid., Ch. II, particularly the long note concerning varieties of intuitionism on pp. 17-20.

[11]Most dramatically, as Sidgwick observed, when people set their sights on ends which will only be realized if at all after they have ceased to exist:

> [M]en have sacrificed all the enjoyments of life, and even life itself, to obtain posthumous fame:

will hardly settle the first question, however, since defenders of hedonistic conceptions of welfare typically use the term 'pleasure' in an extended sense, to refer broadly to something like the "satisfaction" which individuals derive from their lives. It is not obviously implausible to suppose that the amount of pleasure which individuals derive from their lives in this extended sense is exactly proportioned to the strengths of their desires for various things -- e.g., that the amount of satisfaction which seekers of posthumous fame derive from a reasonable certainty of success (or what comes to the same, the substantial dissatisfaction attending a suspicion of failure) is exactly proportioned to the strengths of their desires for posthumous fame -- so that pleasure in the hedonist's sense really does serve as a precise measure of the extent to which people value things.

Even if it is true, however, that under certain conditions pleasure may serve as an accurate indicator of value, it cannot serve in a fully adequate analysis of welfare. There is a deeper issue between preference-based and pleasure-based accounts, residing in the fact that the former possess an objective component which the latter

---

not from any illusory belief that they would be somehow capable of deriving pleasure from it, but from a direct desire of the future admiration of others, and a preference of it to their own pleasure. (The Methods of Ethics, 7th ed., London: MacMillan & Co. Ltd., 1907, pp. 51-2.)

lack.[12] Whether we are speaking of pleasure in the ordinary sense, or whether we have in mind pleasures and pains in the wider sense of satisfactions and dissatisfactions, pleasures are a felt aspect of one's experiences: they may depend upon one's desires and on one's beliefs about the world, but they do not necessarily depend on what the world is like independently of one's psychological state. The satisfaction of desires, on the other hand, does normally[13] depend on features of the world which are independent of one's state of mind. The difference shows up clearly in the possibility of being mistaken about whether one's desires are satisfied. If I keenly desire posthumous fame, e.g., then I may derive a good deal of satisfaction from the false belief that my goal is likely to be met. Thus even if is true that degrees of pleasure or satisfaction are perfectly proportioned to strengths of preference -- i.e., that the amount of pleasure I take in thinking that my ends are satisfied is a precise measure of the extent to which I value them -- pleasure and the satisfaction of preference are still fundamentally different things, and it is possible to promote one without promoting the other.

The mere fact that pleasure and preference satisfac-

---

[12]Cf. Griffin, ibid., as well as Sec. 2 of his article "Modern Utilitarianism", Revue Internationale de Philosophie 36 (1982), pp. 331-375.

[13]I.e., except in cases where what one desires is specifically a certain state of mind.

tion are different things doesn't show that a preference-based conception of welfare is better suited to the utilitarian's purposes than a pleasure-based one, however. Where pleasure and preference satisfaction do in fact part company, isn't it the former rather than the latter that one should attend to in promoting welfare? After all, people will be happy enough if they think that their desires are satisfied, regardless of whether they actually are or not. Isn't the utilitarian's main concern to promote happiness in this sense?

The answer must be "no", at any rate if the utilitarian is to remain faithful to the convictions mentioned above. Utilitarians seek to ground moral value in what individuals actually value, and the things which people value are typically distinct from the satisfaction they take in thinking them realized. If one were forced to choose between having some pressing desire satisfied but falsely believing that it was unsatisfied, or else not having the desire satisfied but falsely believing that it was, I hazard to think that in some instances at least some people would choose the former alternative. To override individuals' preferences here and take the view that the latter alternative is really the <u>morally</u> preferable one is in fact to abandon the search for a ground of moral value, and to adopt one more intuitionistic standard among the rest: 'Pleasure is morally valuable (never mind whether

anyone actually values it)'. Hedonistic intuitionism of this sort is surely less objectionable in certain respects than many other intuitionisms, but it is none the better from a foundational point of view for all that. If we are really to divest ethics of intuitionistic precepts and ground moral value in the empirical world, then it is to individuals' preferences that we must look rather than to what gives them pleasure.[14]

---

[14]Similar remarks apply to more "objectivist" construals of welfare. Thomas Schwartz writes:

> The subjectivist conception [of welfare; i.e. the conception which identifies a person's welfare with the satisfaction of her preferences, tastes, or desires] is preposterous. Human welfare, ordinarily so-called, is nothing like what the subjectivist says it is. Neither can the subjectivist surrogate bear the normative burden of the genuine article. ("Human Welfare: What It Is Not" in Harlan B. Miller and William H. Williams, eds., The Limits of Utilitarianism, Minneapolis: University of Minnesota Press, 1982, p. 195.)

After attacking subjectivist views Schwartz goes on to sketch his own account, which more-or-less identifies welfare with the satisfaction of needs, those things necessary for individuals to function (or function well) as human beings (he doesn't use the term 'needs', however). In effect he invites us to abandon the search for a ground of moral value and adopt an intuitionistic standard not unlike that of the hedonistic intuitionist: 'Human welfare (the real stuff, that is) is morally valuable, never mind whether anyone values it.' The appropriate utilitarian response is to give Schwartz the term 'welfare', reformulate utilitarianism directly in terms of what individuals value, and then observe that on the utilitarian view of things so-called "human welfare ordinarily so-called" has moral significance just to the extent that individuals value it. Whether utilitarian value so-construed is incapable of bearing the "normative burden" which Schwartz thinks can be borne by his intuitionistic surrogate is a question which cannot be fruitfully addressed at this early stage of our inquiry; let me simply note here that as we proceed we will discover reasons for thinking that he is

I shall accordingly be complying in what follows with the view that welfare is the satisfaction of preferences, broadly construed. There have been some important dissenters from the preference-based view,[15] and we will touch on some of their reasons for thinking that preference-based accounts of welfare are inadequate in Section 5 below. In the meantime, it is important not to overestimate the differences of opinion between the so-called classical hedonists and the majority of contemporary theorists on this score. While it is true that Bentham and Mill maintained that utilitarianism's sole concern is with promoting pleasure over pain, they also subscribed to the doctrine that Sidgwick later called "psychological hedonism", the view that people never desire anything as an end in itself except pleasure or the avoidance of pain, and that anything else which they might desire is only desired as a means to increasing their balance of pleasure over pain. Now, psychological hedonism is not a very plausible doctrine unless 'pleasure' is given an extremely wide reading (which of course Bentham and Mill did), and in any case seems falsified by the theoretical possibility of individuals choosing to have their desires satisfied in the knowledge that they will falsely believe them to be

---

mistaken in this regard.

[15]Notably, Richard Brandt; see "Two Concepts of Utility" in Miller and Williams, The Limits of Utilitarianism, pp. 169-85, as well as A Theory of the Good and the Right (Oxford: Clarendon Press, 1979), Ch. XIII.

unsatisfied. The merits of the doctrine aside, however, there are two things to notice here. The first is that Bentham's and Mill's attachment to psychological hedonism makes it difficult if not impossible to determine the extent of their commitment to a pleasure-based account of welfare on the crucial issue of whether one should promote pleasure at the expense of satisfying desires, since if individuals really desire nothing but pleasure, then there is no difference between promoting their pleasure and promoting the satisfaction of their desires.

The second thing to notice is that Bentham and Mill thought it <u>important</u> to explicitly state and defend psychological hedonism. They did not simply announce that pleasures were good and pains bad from a moral point of view; promoting pleasure over pain was a good thing for them precisely <u>because</u> that is what (on their view) people hold to be intrinsically valuable. Given their attacks on intuitionism in its various guises, there is no reason to suppose that they would have opted for hedonistic intuitionism at the crucial point. Indeed, the opening lines of Mill's "proof" of the principle of utility are most plausibly interpreted as an explicit statement of the foundational convictions mentioned above, and commit him directly to a preference-based conception of welfare for

the reasons given.[16] Hence the adoption of a preference-based account is perhaps best viewed as a needed clarification or generalization of the classical utilitarians'

---

[16] "The only proof capable of being given that an object is visible, is that people actually see it. The only proof that a sound is audible, is that people hear it: and so of the other sources of our experience. In like manner, I apprehend, the sole evidence it is possible to produce that anything is desirable, is that people do actually desire it." (J.S. Mill, Utilitarianism, On Liberty, and Considerations on Representative Government, edited by H.B. Acton, London: J.M Dent & Sons, 1972, p. 33.)

This passage has attracted a good deal of critical attention through the years, but most of it has failed to locate Mill's remarks in the context of his empiricist methodology, and in light of his hostility towards intuitionistic principles. Mill is not guilty (as G.E. Moore suggested) of a simple equivocation between something's being desirable in the sense of 'capable of being desired' and in the sense of its being something which 'ought to be desired'. Nor is he claiming that the mere fact that someone desires something shows that she or anyone else ought to desire it; he is not, that is, committing the so-called "naturalistic fallacy". (Cf. his initial remarks concerning the "proof", p. 4: "Whatever can be proved to be good [in the ordinary sense of 'proof'], must be so by being shown to be a means to something admitted to be good without proof.")

The real issue here, the issue between utilitarians on the one hand and intuitionists of various stripes on the other, concerns what things are morally valuable, and Mill is insisting that this contentious issue should be settled by reference to what people actually value. The sole evidence we should countenance for something's being morally desirable, according to Mill, is that people do desire it. The comparison with visibility and audibility is not at all inapposite, given general strictures regarding settling contentious claims by reference to agreed-upon empirical reality. If someone claims that there are ghosts, I may reasonably ask to be shown the evidence. Similarly, if someone claims that there is moral value in the world, I may reasonably ask to be shown the evidence; and the only remotely plausible candidates to be held up as evidence here -- the only things which may be "admitted to be good without proof" -- are the things which people do as a matter of fact value.

views, rather than a substantial revision of them.[17]

2. History of the Problem

However important the distinction between pleasure-based and preference-based conceptions of welfare for an adequate formulation of utilitarianism, little in the debate over interpersonal comparisons actually turns on adopting one sort of account rather than the other. The key difference between the two conceptions lies in the objective or mind-independent implications of satisfying someone's preferences, and in this respect I have argued that preference-based accounts provide a better expression of fundamental utilitarian convictions than pleasure-based accounts do. In attempting to rank actions or policies with respect to the extent that they promote aggregate welfare, however, what we must focus on are individuals' prospective welfares, as determined by the strengths of their preferences for various outcomes. Inasmuch as both degrees of pleasure or satisfaction and degrees of preference are psychological magnitudes, the measurement and interpersonal comparison of either poses similar difficulties. Hence for purposes of the discussion at hand we may

---

[17]On this score as well as others Sidgwick cannot be classified as a classical utilitarian. Having carefully presented his case against psychological hedonism, he went on to explicitly adopt hedonistic intuitionism as his account of moral value (see e.g. Methods of Ethics, Book III, Chapter XIV).

conveniently gloss over the distinction between the two types of account.

A brief sketch of the history of the problem of interpersonal comparisons will help to set the stage for the arguments to follow in the remainder of this chapter and the next. Utilitarians and utilitarian-minded economists of the 18th and early 19th centuries apparently recognized no theoretical obstacles standing in the way of comparing different individuals' welfares. Early scepticism about interpersonal comparisons was due mainly to economists in the latter part of the 19th century, and went hand in hand with the recognition that utility comparisons are not required for explaining market phenomena. W. Stanley Jevons was one of the first to explicitly raise doubts about the possibility of performing objective welfare comparisons; with reference to the theory of exchange developed in his Theory of Political Economy he wrote:

> The reader will find...that there is never, in any single instance, an attempt made to compare the amount of feeling in one mind with that in another. I see no means by which such comparison can be accomplished. The susceptibility of one mind may, for what we know, be a thousand times greater than that of another. But, provided that the susceptibility was different in a like ratio in all directions, we should never be able to discover the difference. Every mind is...inscrutable to every other mind, and no common denominator seems to be possible. But even if we could compare the feelings of different minds, we should not need to do so; for....the motive in one mind is weighed only against other motives in the same mind, never against the motives in other

> minds....Hence the weighing of motives must always be confined to the bosom of the individual.[18]

There is a good deal of insight imbedded in these remarks, I think, not simply in Jevons' perceptive formulation of the problem of interpersonal comparisons (viz., that "provided that the susceptibility was different in a like ratio in all directions, we should never be able to discover the difference"), but also in his observation that individual behaviour is inevitably determined by the weighing of motives within a single mind. The latter observation provides an important clue towards resolving the problem of interpersonal comparisons, one to which we shall return in Chapter 3.

One thing which is notably absent from Jevons' remarks is an acknowledgement of the role that interpersonal comparisons play in ordinary thought and speech. While it may be true that market behaviour can be explained solely by reference to the weighing of motives "confined to the bosom of the individual", it is also true that we often do attempt to make judgments concerning the relative strengths of different individuals' preferences, and that we do so

---

[18]W. Stanley Jevons, The Theory of Political Economy, 4th ed. (London: MacMillan and Co., Ltd., 1911; 1st ed.: 1871), p. 14. Note that although Jevons' avowed aim was to develop a theory of exchange "entirely based on a calculus of pleasure and pain" (p. 23), his remarks here and elsewhere are concerned primarily with the weighing of individuals' "motives", i.e. the strengths of their preferences or desires. This ambiguity is typical of the late 19th and early 20th century literature.

with some degree of confidence. When we are reasonably familiar with people, we seem to have little difficulty formulating judgments about interpersonal utilities with respect to fairly finely individuated states of affairs. I have no doubt, e.g., that on typical occasions certain of my friends would prefer a glass of single malt scotch to a blend a good deal more than others; or that my wife's preference for watching a movie rather than a ball game is often stronger than my preference for watching the game. When we are not very familiar with people we are not of course in a position to make even rough comparisons of this sort with any degree of confidence, but we have little trouble in arriving at more coarse-grained comparisons. Philosophical doubts aside, is there anyone who really supposes that having a color rather than a B&W television means as much to a typical middle-class North American as having a few decent meals a day means to a starving person?

The fact that substantial intersubjective agreement does exist with respect to judgments of this sort lends prima facie support to the case for objective interpersonal comparisons. The support is limited, however. While our ordinary judgments of interpersonal welfare certainly seem to be objective enough, it is far from easy to ascertain their exact ground. Judgments concerning the welfares of people we are familiar with presumably depend in some measure on observations of their behaviour in various

circumstances, including their verbal behaviour and other expressions of their likes and dislikes. More coarse-grained judgments concerning people we are not familiar with perhaps depend on some broad generalizations about the nature of human beings and the environments in which they are situated. Just what more is involved in performing welfare comparisons beyond these very vague indications is difficult to say, however. Interpersonal comparisons have tenaciously resisted satisfactory analysis, and without some fuller accounting of the facts underlying them, the possibility remains open that our everyday judgments are largely groundless, no matter how confidently advanced, and that utilitarianism is for the most part an empty and confused doctrine.

That being said, we should recognize that the prima facie support for interpersonal comparisons generated by widespread intersubjective agreement places a certain burden on the sceptic. If it is really true that there is no way of objectively comparing the strengths of different individuals' preferences, then a whole class of judgments which we are inclined to advance with a good deal of confidence apparently rest on some sort of colossal mistake. The sceptic's burden is to explain the nature of this mistake, why it is made in common by so many people, and why it is such an easy one to overlook; without an explanation of this sort, scepticism about interpersonal

comparisons remains merely negative, and not entirely convincing. (I am not suggesting that Jevons was remiss in failing to provide such an explanation; he was primarily concerned to point out that utility comparisons are not required for explaining market behaviour, and within the context of that discussion his negative remarks are perfectly in order.)

Jevons' doubts about the possibility of performing interpersonal comparisons did not extend to the measurement of single individuals' utilities. While expressing some reservations about our ability to accurately determine the strengths of motives differing widely in extent, he supposed that it was possible for a person to introspectively determine differences in strength of motivation with a fair degree of precision when the strengths do not differ too greatly.[19] Edgeworth took exception to this double standard, as he saw it,[20] and appealed to the notion of a "minimum sensible" or just noticeable difference in utility

---

[19] Ibid., pp. 12-13.

[20] "There is, no doubt, much difficulty here, and the risen science is still obscured by clouds; and hedonism may still be in the state of heat or electricity before they became exact sciences, as described by Professor Jevons. Let us, however, following in his footsteps, endeavour to gain as clear a view as may be. At least it is hoped that we may sight an argumentum ad hominem, an argument to the man who (with Professor Jevons), admitting mathematical reasoning about self-regarding pleasures, denies the possibility of mathematically comparing different persons' pleasures." (F.Y. Edgeworth, Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences, London: C. Kegan Paul & Co., 1881, p. 98.)

as the appropriate unit both for measuring a given individual's welfare and for comparing welfares across persons: "Just-perceivable increments of pleasure, of all pleasures for all persons, are equatable".[21]

Edgeworth's method of just noticeable differences has not received much support as a means of measuring and comparing welfare, and with good reason. The primary difficulty is that the method provides no non-question-begging way of identifying just noticeable differences in pleasure or preference, as distinct from just noticeable differences in the objects of pleasure or preference.[22] Suppose that I happen to think that 1 teaspoon of sugar in my coffee is just about the right amount, so that I prefer 1 teaspoon to 2 teaspoons, and to 1-1/2 teaspoons, 1-1/4

---

[21] Ibid., p. 60. See also pp. 7-8, as well as Appendix III.

[22] Edgeworth must be able to draw this distinction, on pain of being open to the charge that his method would grant more weight in calculations of aggregate utility to some people's preferences merely in virtue of their being more capable of discriminating various features of the physical world. (Cf. Sen, Collective Choice and Social Welfare, pp. 94-5, and Rawls, A Theory of Justice, p. 321-2. On the assumption that the distinction in question can be drawn, what Sen and Rawls appear to regard as the major objection to the method of just noticeable differences -- that it would equate differences which different people feel differently about -- seems itself question-begging, since the method is precisely supposed to define what it means for people to feel the same or differently in the requisite sense. For critiques of more recent attempts to ground judgments of interpersonal welfare on just noticeable differences in utility, see Kenneth J. Arrow, Social Choice and Individual Values, 2nd ed., New Haven: Yale University Press, 1963, pp. 115-18; and Jerome Rothenberg, The Measurement of Social Welfare, Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1961, Ch. 7 & 8.)

teaspoons, and 1-1/8 teaspoons. At some point, if not here then fairly soon, my powers of discrimination begin to run out. Suppose that I "barely" prefer 1 to 1-1/16 teaspoons, but that I am incapable of discriminating between, and hence am indifferent to, amounts of sugar below this threshold. Does this tell us anything of interest about the strengths of my preferences, as opposed to the discriminatory powers of my taste buds?

It is not clear that it does. In order to preserve the distinction between minimal degrees of pleasure or preference and just noticeable differences in the objects of pleasure or preference, Edgeworth must allow that my current preference for 1 over 1-1/16 teaspoons might not be a true "bare preference" of the requisite sort, and hence that if my discriminatory powers increased I might come to prefer, say, 1 teaspoon to 1-1/32 teaspoons of sugar in my coffee, without the strength of my original preference for 1 over 2 teaspoons thereby increasing. But this in turn raises the question of whether Edgeworth's "minimum sensible" exists at all. I can think of no principled reason for supposing that if my power to discriminate amounts of sugar were to increase indefinitely, my discriminations of pleasure or preference might not follow suit, while the strength of my preference for 1 over 2

teaspoons remained intact.[23] Of course, my power to discriminate amounts of sugar could not proceed beyond the level of detecting the presence or absence of individual molecules. But this does not affect the point that a "bare" preference for something depends greatly on, and may be completely exhausted by, our capacity to discriminate features of objects (whether because of our own limitations or because of limits on the divisibility of the objects themselves), and that there may be no such thing as an atom of preference or pleasure per se. And if there is no way of demonstrating that atoms of preference or pleasure exist, let alone of identifying them for a given individual, then so much the worse for counting them up as a means of comparing different individuals' utilities.

To return to our story: Early scepticism about interpersonal comparisons coincided with the recognition that welfare comparisons are not required for explaining market behaviour. A new twist was added to the debate

---

[23] It may seem implausible to suppose that the strength of my original preference for 1 over 2 teaspoons would remain constant under such circumstances. If the discriminatory powers of my taste buds were really so finely honed as the example suggests, wouldn't a cup of coffee with 2 teaspoons of sugar in it come to seem intolerably sweet, rather than just somewhat too sweet as it does now? Perhaps so. Keep in mind, however, that the objection here is a principled one: in order to maintain the distinction between minimum sensible differences in pleasure or preference and the objects of pleasure or preference, Edgeworth must allow that it is at least theoretically possible that my discriminatory powers might increase indefinitely while the strength of my original preference remained intact.

around the turn of the century when Pareto noticed that, given certain idealizing assumptions about the nature of markets and economic goods, market equilibria can be explained on the basis of "indifference curves" representing individuals' purely ordinal preferences for various bundles of goods. Indifference curves had been introduced into the literature by Edgeworth,[24] but he had derived them from the assumption of "cardinal" or measurable utility. Pareto on the other hand adopted the strategy of taking the indifference curves themselves as basic, without assuming that utility (or "ophelimity" as he called it) is a measurable quantity underlying the ordinal representations.[25] Pareto's project was carried to completion in an influential article by Hicks and Allen, which among other things replaced the suspect notion of diminishing marginal utility (which Pareto had relied on heavily in the expository portions of his work) with the notion of "increasing marginal rates of substitution".[26]

---

[24]Ibid., pp. 21f.

[25] Vilfredo Pareto, Manual of Political Economy, translated from the French edition of 1927 by Ann S. Schwier (New York: Augustus M. Kelley, 1971), pp. 110-13, 118f., 191f., 391f.

[26]J.R. Hicks and R.G.D. Allen, "A Reconsideration of the Theory of Value", Economica Vol. 1, No. 1 (February 1934), pp. 52-76 and Vol. 1, No. 2 (May 1934), pp. 196-219. It should be noted that Hicks' and Allen's increasing marginal rate of substitution captures only part of the notion of diminishing marginal utility as Pareto conceived of it. Indifference curves are standardly drawn convex to the origin, reflecting the fact that as individuals give up successive increments of a good represented on the x axis, they must be compensated with progressively greater amounts

It remained for Lionel Robbins to acknowledge that we often do make judgments purportedly comparing different individuals' welfares, and to attempt to provide some explanation of them. Since market behaviour can apparently be explained on the basis of individuals' ordinal rankings

---

of the good represented on the Y axis. Pareto also supposed, however, that it is possible for individuals to make rough judgments concerning the amount of utility they receive from increments of a single good, quantities of all other goods remaining fixed; and that gains in utility in this sense generally decrease with successive increments of the good in question. He represented this aspect of diminishing marginal utility by the shape of his "hill of ophelimity": If we think of a standard indifference map as a topographical map with the indifference curves representing utility-elevations rising from the origin, then according to Pareto the resulting hill rises sharply at first, and then gradually levels off as the individual moves to higher indifference curves. Increasing marginal rates of substitution determine only the convex shape of the indifference curves; the notion of diminishing marginal utility as reflected in the shape of the hill of ophelimity is dismissed by thorough-going ordinalists as irrelevant at best.

(Cf. Hicks and Allen, p. 57. Hicks misleadingly describes the component in Pareto's concept of diminishing marginal utility which is not captured by increasing marginal rates of substitution as follows: "that the marginal rate of substitution will increase, not only when Y is substituted for X, but also when the supply of Y is increased without any reduction in the supply of X." This appears to be a misinterpretation of Pareto's view, perhaps fostered by Hicks' failing to explicitly notice that the former's notion of "elementary ophelimity", corresponding to the classical concept of marginal utility, concerns movement between indifference curves; whereas the Hicks-Allen replacement for the classical concept (i.e., the rate of substitution of Y for X) is exclusively concerned with the slope of a single indifference curve. In any case, as Hicks later emphasizes and as Pareto was well aware, it is perfectly conceivable that for some fixed values of X the marginal rate of substitution of Y for X may remain constant or decrease when the supply of Y is increased; the concept of diminishing marginal utility as reflected in the shape of the hill of ophelimity is quite independent of rates of substitution.)

of bundles of goods, without supposing that their preferences have magnitudes or intensities at all, Robbins concluded in all narrowly-focused positivistic reasonableness that utility is not measurable for the individual, and a fortiori not comparable across individuals. Thus ordinary judgments of interpersonal welfare cannot rest on objective facts, but must be understood as a kind of disguised value judgment expressing our views about how goods ought to be distributed among individuals.[27]

Though Robbins did not explicitly say so, his subtext made it clear that intersubjective agreement in judgments of interpersonal welfare was to be explained in virtue of the fact that people in a given culture share substantially the same values. He appears to have been particularly impressed with a story related by Sir Henry Maine concerning a Brahmin who had been apprised of some of the consequences of Benthamite utilitarianism b a government official:

> "But that," said the Brahmin, "cannot possibly be right. I am ten times as capable of happiness as that untouchable over there." I [Robbins] had no sympathy with the Brahmin. But I could not escape the conviction that, if I chose to regard men as equally capable of satisfaction and he to regard them as differing according to a hierarchical schedule, the difference between us was not

---

[27]See L.C. Robbins, An Essay on the Nature and Significance of Economic Science, 2nd ed., revised and extended (London: MacMillan and Co., 1945), Ch. VI; "Interpersonal Comparisons of Utility: A Comment", Economic Journal 48 (1938), pp. 635-41; and "Robertson on Utility and Scope", Economica 20 (February 1953), pp. 99-111.

> one which could be resolved by the same methods of demonstration as were available in other fields of social judgement.[28]

Robbins' claims about the essentially normative character of interpersonal comparisons attracted a good deal of critical attention from economists who took him to be arguing that they ought to refrain from offering policy recommendations altogether. But in fact he had urged no such thing. Disagreement on judgments of interpersonal welfare cannot be resolved in a "purely scientific manner", Robbins thought, and hence must rest on normative beliefs. Some value judgments are eminently more defensible than others, however, and there is no reason why economists should fail to take a stand on these "philosophical" issues. Much better that they should explicitly take a stand on the issues, rather than try to pass off their egalitarian views as factual claims belonging to the sphere of economics proper.

By the 1930's, then, economists had largely occammed cardinal or measurable utility out of existence, and with it any hope of arriving at factual comparisons of individuals' welfares. If interpersonal comparisons were to be defended at all, they would have to be defended on explicitly normative grounds.[29] This happy coalescence of

---

[28] "Interpersonal Comparisons of Utility", p. 636.

[29] In "A Reformulation of Certain Aspects of Welfare Economics" (Quarterly Journal of Economics 52, 1938, pp. 310-34) Abram Bergson developed a general framework designed to isolate the value judgments which he presumed

opinion was short-lived, however. The notion of cardinal utility was to some extent rehabilitated in 1944, when von Neumann and Morgenstern announced that (to borrow Daniel Ellsberg's phrase)[30] they had "succeeded in synthesizing 'measurable utility'" via an examination of individuals' choices involving risky prospects.[31] The heated debate

---

to be involved in analyses of social welfare. Bergson's article proved to be seminal in laying the groundwork for social choice theory as developed by Arrow and others in the 1950's. Cf. also Oscar Lange, "The Foundations of Welfare Economics", Econometrica 10 (1942), pp. 215-28, where assumptions concerning the measurability and comparability of utility were similarly eschewed in favor of explicitly postulating a "social valuation of the importance of individuals".

It should be noted that not all economists of the period were content to relegate judgments of social welfare (beyond those supported by the Pareto criterion; see p. 13 above) to the realm of the purely normative. Apart from the odd economist who remained unconvinced of the non-measurability and thus non-comparability of utility, some were attracted to the idea that aggregate welfare could be said to have increased if, after a movement from one social state to another, it was possible to compensate the "losers" through redistribution so that they would be no worse off than they were in the former state; see Nicholas Kaldor, "Welfare Propositions and Interpersonal Comparisons of Utility", The Economic Journal 49 (1939), pp. 549-52, and Tibor Scitovsky, "A Note on Welfare Propositions in Economics", The Review of Economic Studies 9 (1941), pp. 77-88. Unfortunately, the compensation test as originally devised by Kaldor is inconsistent, and Scitovsky's attempt to patch it up yields a non-transitive social preference relation; for discussion see Sen, Collective Choice and Social Welfare, Ch. 2*. (Kaldor's and Scitovsky's articles are both reprinted in Kenneth J. Arrow and Tibor Scitovsky, eds., Readings in Welfare Economics, London: George Allen and Unwin Ltd., 1969, as are the articles of Bergson and Lange mentioned above.)

[30] "Classic and Current Notions of 'Measurable Utility'", The Economic Journal 64 (1954), p. 528.

[31] See John von Neumann and Oskar Morgenstern, Theory of Games and Economic Behaviour, 2nd. ed. (Princeton: Princeton University Press, 1947), pp. 15-31. Frank Ramsey had developed an axiomatization of utility similar to the

which ensued between ordinalists and cardinalists over the status of measurable utility and its role in economic theory is an interesting one in its own right, but we won't pause to consider all of the gory details. The point of immediate relevance is that the vN.M. procedure yields at best an interval measure of utility (of which, more in the next section) for each individual, and hence leaves open the question of whether objective interpersonal comparisons are possible. "We re-emphasize," they wrote, "that we are considering only utilities experienced by one person. These considerations do not imply anything concerning the comparisons of the utilities belonging to different individuals."[32]

So far as the problem of interpersonal comparisons is concerned, then, the net effect of the vN.M. synthesis of cardinal utility was to return the debate to its origins, though with a somewhat sharper understanding of some of the issues involved. Where Jevons had maintained simply that

---

vN.M. one some 15 years earlier, the primary difference being that Ramsey's procedure simultaneously defines subjective probabilities and utilities. (See "Truth and Probability" in Foundations: Essays in Philosophy, Logic, Mathematics and Economics, Atlantic Highlands, N.J.: Humanities Press Inc., 1978, pp. 58-100; originally published in The Foundations of Mathematics and Other Logical Essays, Routledge & Kegan Paul, 1931.) Ramsey's work was overlooked by economists, presumably because his main concern was to axiomatize subjective probability rather than utility per se. Von Neumann and Morgenstern briefly mention the possibility of axiomatizing probability and utility together, without mentioning Ramsey's construction, in a footnote on p. 19.

[32]Ibid., p. 29.

utility was measurable for the individual but not comparable across individuals, the improved understanding of measurement procedures which was precipitated by the work of von Neumann and Morgenstern now made it possible to speak directly in terms of interval representations of utility for each individual, and to wonder what more was required beyond such representations in order to secure interpersonal comparability. Still, the crux of the issue remained: Assuming that we have in hand the requisite interval scales, is there any means of detecting whether one person's preferences are on balance a thousand times stronger than another's, supposing that their preferences differ "in a like ratio in all directions"? Or must we finally agree with Robbins that our ordinary judgments of interpersonal welfare are really disguised value judgments, erroneously advanced as factual claims?

Two sorts of proposal have been offered for meeting Jevons' and Robbins' sceptical challenges. The most common response has been that we can and do make objective interpersonal comparisons on the basis of differences in individuals' behaviour. More recently, Richard Brandt has suggested that welfare comparisons can be made on the basis of differences in individuals' physiologies. Following some clarificatory remarks on the nature of interval utility functions, we examine these proposals for performing interpersonal comparisons in turn.

3. Interval Utilities

In the remainder of this chapter and the ones following I shall be assuming that individuals' preferences over any set of alternatives may be arranged on interval scales, in accordance with the vN.M. procedure for constructing personal utility functions. I shall further assume that such rankings provide grounds for talking about the relative magnitudes or intensities of an individual's preferences: we can say that she prefers x to y twice as much as she prefers z to w, one third as much as she prefers u to v, and so on. Interval rankings do not of course license stronger claims to the effect that a given preference has an "absolute" magnitude.[33]

The supposition that vN.M. utility functions provide a basis for speaking of the relative strengths of an individual's preferences is not beyond reproach.[34] Assuming that preferences can meaningfully be understood to have

---

[33]The reader is cautioned not to interpret the term 'absolute' as marking some mysterious metaphysical distinction (as though preferences might have their magnitudes stamped on them by God in indelible ink, independently of how we measure them). It simply means "not relative", in this case not relative to the strengths of an individual's other preferences. In a similar vein, if someone were to ask me how tall my youngest brother is and I replied that he is 1.04 times as tall as my oldest brother, the questioner might well respond that she wanted to know his height in "absolute" terms, not his height relative to someone else. Presumably what she wants to know is his height relative to a standard meter or yardstick.

[34]Note however that von Neumann and Morgenstern did maintain the supposition themselves (ibid., p. 18), contrary to what some commentators appear to suggest.

magnitudes at all, more than a few theorists have wondered whether the vN.M. procedure can be appropriately construed as measuring degree of preference for risk-free outcomes, given that it relies on purely ordinal information concerning preferences for risky prospects. A quick sketch of the vN.M. measurement procedure will help to clarify the nature of the objection.

Suppose we know that A prefers x to y and y to z. How can we get beyond this mere ordering of outcomes and find out, e.g., whether in A's estimation y is almost as good as x, or whether she holds y to be just a little bit better than z? The core intuition underlying risk-based measures of utility is that we can discover the extent of someone's preferences for various outcomes by finding out what risks they are prepared to take in order to satisfy them. Thus if we give A the outcome y for certain, we can discover something about the strengths of her preferences by offering her various "lotteries" over x and z in exchange for y and observing which ones she will accept and refuse. (A lottery over x and z is a gamble which yields prize x with probability p and prize z with probability 1-p, e.g. a 20% chance of x and a 80% chance of z; p may be equal to 1 in the case of a "lottery" offering one prize with certainty.) The idea is roughly that if in A's estimation y is almost as good as x, she will be reluctant to give up the certainty of the former unless she is offered a fairly good

chance of getting the latter (equivalently, unless the risk of ending up with z is quite low); whereas if she holds y to be just a little bit better than z, she will accept a lottery which yields a fairly low probability of x, since if she ends up with z she won't have lost very much anyway.

Depending on where A subjectively locates y with respect to x and z, it is reasonable to suppose that there will be one and only one lottery over the latter two outcomes which she holds to be indifferent to the certainty of y. What von Neumann and Morgenstern demonstrated is that, if A's preferences satisfy certain plausible-looking axioms, there is a real-valued function U assigning numbers ("utilities") to non-risky outcomes in such a fashion that any lottery over the outcomes is preferred to any other if and only if the "expected utility" of the former -- the utilities of the prizes in the lottery, weighted by their probabilities -- is greater than that of the latter.[35] More concretely: Suppose that A's preferences satisfy the requisite axioms and that she is indifferent between the

---

[35]The vN.M. axioms are a bit involved, and simpler systems have been proposed. The substantive requirements are roughly that an individuals' preferences must weakly order all alternatives, including simple and compound lotteries (i.e., lotteries in which one or both of the prizes is itself a lottery); that if one alternative is preferred to another and that to a third, there is just one lottery over the first and third alternatives which is indifferent to the second; and some version of the "sure-thing" principle, e.g. one alternative is preferred to another if and only if any lottery over the former and some third alternative is preferred to the equivalent lottery with the second alternative substituted for the first.

certainty of y and a lottery offering a 70% chance of x and a 30% chance of z. Then if we arbitrarily assign the value 1 to outcome x and 0 to outcome z, in order to explain A's choices in terms of the hypothesis that she maximizes her expected utility we must assign a value of .7 to y (since 70%x1 + 30%x0 = .7).

Since the endpoints on the partial utility scale we have just constructed were arbitrarily chosen, it would clearly not do to maintain, say, that A's preference for x over y has a magnitude of .3. We might just as easily have chosen the values 3 and 0 for x and z respectively, in which case y would have received a value of 2.1 and the difference between U(x) and U(y) would be .9; or 100 and -5 (in which case U(x) - U(y) = 100 - 68.5 = 31.5).

Given that the values for x and z were arbitrarily chosen, the resulting utility function U(x)=1, U(y)=.7, U(z)=0 is said to be unique up to increasing linear transformation. More generally, if U is a vN.M. utility function representing an individual's preferences over some set of outcomes, then so is any U' = aU + b (with a>0). (Another way of expressing uniqueness up to linear transformation is to say that the measurement procedure yields a scale which is unique up to the selection of a "zero-point" and "unit of measurement"; selection of the positive constant a fixes the unit, while selection of b fixes the

origin.)[36]

Thus depending on how we choose the endpoints, any one of an infinite family of scales, each a positive linear transformation of the others, will adequately represent the facts about A's preferences concerning x, y, and z which we used to construct our original scale. Numerical differences between given points on these scales clearly vary from one scale to another. There is however something which is invariant across all the scales: namely, the ratios of the intervals between outcomes. I.e., the ratio of the interval between x and y to the interval between y and z (.3/.7) is constant across the entire family. For this reason vN.M. utility scales are called "interval scales", and the procedure is said to generate an "interval measure" of utility. (I shall sometimes refer to the result of the vN.M. procedure simply as an "interval ranking" of outcomes.)

We are now in a position to appreciate the force of the objection mentioned above. The fact that ratios of intervals are invariant with respect to linear transformations may plausibly be viewed as providing grounds for thinking that vN.M. utility functions represent the relative magnitudes of individuals' preferences for risk-

---

[36]The standard reference for these notions is David H. Krantz, R. Duncan Luce, Patrick Suppes and Amos Tversky, Foundations of Measurement, Vol. I (New York: Academic Press, Inc., 1971).

free outcomes -- e.g., for holding that the strength of A's preference for x over y is 3/7 the strength of her preference for y over z. The objection in question insists that there are no such grounds, however, since vN.M. utility scales are constructed on the basis of information about individuals' choices among risky prospects. The main difficulty is that the procedure does not differentiate between individuals' attitudes towards outcomes and their attitudes towards risk per se. Suppose that A is somewhat "risk-averse", in the sense that she is generally disinclined to give up a sure thing in exchange for a chance of getting something better. Then in the example above, other things equal we would have had to offer her a somewhat greater chance of getting x in order to prompt her to give up the certainty of y. Thus the fact that she is indifferent between y and a 70/30 lottery over x and z cannot be taken to indicate that her preference for x over y is 3/7 as strong as her preference for y over z, since we have no way of determining whether she is risk-averse or not, and if she is then her "true" estimation of the value of y in relation to the certainty of x and the certainty of z would be somewhat lower than our numbers suggest.

This is a real difficulty, one which has not as yet been satisfactorily resolved if indeed it is resolvable at all. I propose to ignore the difficulty here, by assuming that individuals whose utilities are to be compared are

completely neutral w.r.t. risk. As will be clear in the sequel, I do not in any case think that risk-based measures of utility have much of a role to play in real-world applications of the utilitarian calculus. Our primary motivation for assuming that vN.M. utility functions accurately depict the relative strengths of individuals' preferences is that such utility functions are often taken as the starting point in contemporary discussions of interpersonal comparisons; e.g. in Harsanyi's proposals for performing direct interpersonal comparisons, to be discussed in the next section. For purposes of exploring how direct utility comparisons fail, it is important to grant their advocates as much as possible: namely, that we may in principle have access to exhaustive interval rankings representing the true relative strengths of individuals' preferences.[37]

---

[37]It should be mentioned that non-risk-based axiomatizations of preference intensity have been developed; see e.g. Peter C. Fishburn, Utility Theory for Decision Making (New York: John Wiley & Sons, 1970), Ch. 6, and Krantz et. al., Foundations of Measurement, Ch. 4. The main defect of such axiomatizations is that they do not lend themselves well to intuitive "operationalization", whereas the vN.M. construction is motivated from the outset by a concrete and easily understood measurement procedure. Those who object to our provisional assumption of risk-neutrality may, however, replace references to vN.M. utility functions with references to interval utility functions derived from some non-risk-based procedure; the critical assumption in what follows is only that the relative strengths of an individual's preferences are accurately represented by some interval scale. Staunch ordinalists who remain unmoved by the considerations raised immediately below may follow along with the argument in hypothetical mode: "If individuals' preferences have magnitudes, representable by

Having acknowledged the difficulty with risk-based measures of utility, we should note that the objection we have been considering is sometimes run together with another objection which is not entirely cogent.[38] Some commentators have argued that the vN.M. procedure provides no information about the relative strengths of an individual's preferences, since the utility scales are constructed on the basis of purely ordinal information about her preferences for risky prospects. This conclusion does not at all follow from the difficulty mentioned above, however. If A is somewhat risk-averse, then her vN.M. utility function will correspondingly misrepresent her true utilities; the true strength of her preference for x over y will be somewhat greater than 3/7 of her preference for y over z. But it does not follow that A's vN.M. utility function tells us absolutely nothing about the relative strengths of her preferences -- a skewed representation is not the same thing as no representation at all.

The present objection does not rest, then, on the fact that the vN.M. procedure relies on preferences regarding

---

interval scales, then...."

[38]See e.g. "Fallacy 3" in R. Duncan Luce and Howard Raiffa, Games and Decisions: Introduction and Critical Survey (New York: John Wiley & Sons, 1957), p. 32; Fishburn, ibid., pp. 81-2; and Kenneth J. Arrow, "Formal Theories of Social Welfare" in Philip P. Wiener, ed., Dictionary of the History of Ideas, Vol. 4 (New York: Charles Scribner's Sons, 1973), pp. 276-84, reprinted in Arrow, Collected Papers Volume 1: Social Choice and Justice, pp. 115-32.

risky prospects; it rather rests on something like scepticism about preference intensities in general. An individual's vN.M. utility function is constructed on the basis of purely ordinal information about her preferences (never mind that the preferences happen to be for risky prospects), and hence is entirely compatible with the supposition that her preferences don't have magnitudes at all.[39]

Scepticism of this sort may be fairly attributed to many if not most practicing economists, I think. And if vN.M. utility functions were all that we had to go by, we would surely have to grant them their point: the vN.M. procedure does not all by itself provide grounds for thinking that preferences have intensities, based as it is on simple choices among lotteries. If we have antecedent grounds for thinking that preferences have magnitudes, however, then it is not implausible to suppose that vN.M. utility functions go some way towards representing them.[40]

---

[39]Cf. Arrow, ibid., p. 120: "[The vN.M. utility measure] is no longer a measure inherently associated with an outcome; instead, the utility function is precisely that which measures the individual's willingness to take risks." The implication here, I take it, is that risky choices must be explained solely on the basis of attitudes towards risk, since preferences themselves don't have intensities which could account for the choices.

[40]See Rothenberg, op. cit., pp. 211-15, for a good discussion of this point. Harsanyi has also stressed the point in various places; see e.g. "Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory", American Political Science Review 59 (1975), pp. 594-606; reprinted in Essays on Ethics, Social Behaviour, and Scientific Explanation (Dordrecht: D.

And surely we do have such grounds, the fact (if it is one) that preference intensities are not required for explaining market equilibria notwithstanding. I prefer maple walnut to vanilla ice cream a good deal less, in my humble and introspectively tainted opinion, than I prefer a steady diet of whole grains to having no food at all. Subjectively speaking, it is precisely this fact of differing strengths of preference which would prompt me to risk far less to get maple walnut rather than vanilla for dessert this evening than I would to get some porridge if I was starving.

General scepticism about whether preferences have magnitudes is not independent of the question of whether welfares are interpersonally comparable. Such scepticism often rests in part, I suspect, on something like Edgeworth's worry that Jevons was maintaining a double standard in holding that welfare is measurable for the individual but not comparable across individuals.[41] If preferences really do have determinate magnitudes, then they should presumably be measurable by some means or other. But if we can in principle measure the strengths of one individual's preferences, then we might reasonably expect to be able to compare the strengths of different individuals' preferences. No-one has yet managed to

---

Reidel, 1976), pp. 37-63.

[41] See note 20 above.

provide a convincing account of how we can do the latter, however; hence we should be sceptical about whether preferences really do have magnitudes. If this is the kind of reasoning which implicitly underlies ordinalism, then the explanation of why direct interpersonal comparisons fail and the account of indirect comparisons to be developed in the next chapter should help to forestall the gloomy conclusion.

4. Direct Behavioural Comparisons

We have assumed, then, that we have access to interval scales which license statements concerning the relative intensities of an individual's preferences: "A's preference for x over y is half as strong as her preference for z over w" and so on. Given the assumption of personal interval rankings, the problem of interpersonal comparisons comes to this: Is there some meaningful interpretation we can attach to statements concerning the relative strengths of different individuals' preferences? That is, is there some clear sense to be given to the hypothesis that A prefers x to y more than, or less than, or to the same extent as B prefers z to w?

It is of course trivially possible, in one sense of 'possible', to arrive at interpersonal comparisons of preference intensity; namely, on the basis of an arbitrarily chosen "unit of comparison". Since the vN.M.

procedure determines a measure of the intensity of each individual's preferences which is unique up to the choice of a zero point and unit of measurement, we need only resolve to treat some preference of A's as being of exactly the same magnitude as some preference of B's, and given their respective personal scales of preference intensity, we will have automatically generated an exhaustive interpersonal ranking of their preferences.[42] But obviously this is not what is being asked for. What we want to know is whether there is a clear objective sense in which individuals may differ in the extent to which they prefer some alternatives to others, one which doesn't rely on arbitrary assumptions concerning units of comparison. Granted that we have in hand interval scales of utility for each individual, are there any facts of the matter which constrain us from adopting any old positive linear transformation of someone's utility function that we like, prior to making judgments of interpersonal welfare?

---

[42]Note that selection of a unit of comparison will secure only interpersonal rankings of preferences, not interpersonal rankings of outcomes. I.e., the selection would allow us to make comparisons of the form "A's preference for x over y is stronger than B's preference for z over w", but not of the form "x yields more utility to A than y does to B". Given the assumption of personal interval rankings, the latter sort of comparison (sometimes referred to as a "level" comparison) is more demanding than the former, since it requires us to settle not only on a unit of comparison but also to fix the zero-points of individuals' utility scales. For more on level comparisons and their relation to indirect interpersonal comparisons, see the final section of Ch. 3 below.

A number of theorists have suggested that we can secure an objective unit of comparison which will get us beyond personal interval rankings by looking to differences in individuals' behaviour. It should be obvious from the outset, however, that no very simple correlation between behaviour and preference intensity will be forthcoming. For one thing, individuals' behaviour can be expected to vary not only with changes in the intensity of their preferences, but also with changes in their beliefs. Thus two people might have exactly similar preferences (with exactly similar strengths, whatever that finally amounts to); but if they had different beliefs about the world, they might well behave differently in similar situations.[43]

We shall not dwell on this difficulty. While an occasional reminder that logical and methodological behaviourism are dead ends may be useful, focusing on the present difficulty would only serve to mask deeper problems involved in the attempt to compare welfares on the basis of behaviour. Thus for purposes of discussing behavioural

---

[43]Needless to say we are not dealing here with "revealed preferences" of the sort championed by Paul Samuelson ("A Note on the Pure Theory of Consumer Behaviour", Economica 5 (1938), pp. 61-71). On Samuelson's approach, changes in consumptive behaviour entail changes in "preference"; the possibility of someone's beliefs changing in a way which would affect her behaviour while her preferences remained intact is definitionally excluded. For a critique of Samuelson's approach see Amartya Sen, "Behaviour and the Concept of Preference", Economica 40 (1973), pp. 241-59; and Stanley Wong, The Foundations of Paul Samuelson's Revealed Preference Theory (London: Routledge & Kegan Paul, 1978).

proposals we introduce a further assumption: individuals whose utilities are to be compared possess all and only true beliefs about the world.[44]

With that complication out of the way, what sorts of behavioural differences do we or should we take to be interpersonally significant indicators of preference intensity? The answer cannot simply be choice behaviour, for as von Neumann and Morgenstern discovered, examination of an individual's hypothetical choices over even infinitely large sets of risky prospects will at best yield an interval measure of utility, which still leaves us with the problem of locating a non-arbitrary unit of comparison.

The usual tack, defended by I.M.D. Little and John Harsanyi among others, is to focus primarily on individuals' expressive behaviour.[45] Their idea is that the

---

[44]This assumption may seem excessively strong, but I am unsure how to make it weaker. As noted below, behavioural proposals for performing interpersonal comparisons rely on the supposition that the level of satisfaction of an individual's preferences is reflected in her expressive behaviour. Hence it is important (for reasons mentioned in Section 1 of this chapter -- namely, that individuals may derive "satisfaction" from falsely believing that their goals are or will be realized) that people whose utilities are to be compared have true beliefs about which of their preferences are satisfied, or are likely to be satisfied. I see no way of insuring this in general, short of assuming that individuals possess exhaustive sets of true beliefs.

[45]I.M.D. Little, A Critique of Welfare Economics (Oxford University Press, 1950); and John C. Harsanyi, "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility", Journal of Political Economy 63 (1955), pp. 309-21. (Page references to the latter are to the version reprinted in Essays on Ethics, Social Behaviour, and Scientific Explanation, pp. 6-23.)

current level of satisfaction of an individual's preferences will typically be reflected in her countenance, her general demeanor, her verbal expressions of satisfaction or dissatisfaction and the like, and that this behaviour may be taken as a rough guide to the intensity of her satisfied preferences. This proposal has the merit of according well with considerations that seem to be relevant to the judgments of welfare that we do make: other things equal, we judge a person to be happy if she says she's happy, if she smiles a lot, if she whistles while she works, and so on. Conversely, we judge a person to be unhappy if she frowns and complains a lot, etc.; as Little put it:

> [I]f we say of a man that he is always miserable, basing our judgement on how he looks and behaves, and how we know we would feel if we looked and behaved like that, and on a wide knowledge of his character gathered by observing his behaviour and words in a variety of situations, and on the opinions of all his friends who similarly know him well, then we would think it was just nonsense to say that he might really be deceiving everyone all the time and be the happiest of men.[46]

To be sure, interpersonal comparisons based on expressive behaviour will be rough and ready at the best of times. But the acknowledged vagueness of the comparisons, and the attendant latitude for disagreement, should not be confused with lack of objectivity. Refusal to accept behaviour as evidence for interpersonal differences in welfare, according to Little and Harsanyi, amounts to refusing to accept

---

[46]Ibid., pp. 56-7.

behaviour as evidence for other minds.[47]

There may be concerns here about the extent to which various features of individuals' psychologies which are independent of their preferences impact on their readiness to express their satisfaction or dissatisfaction with their situation. Let us grant that such concerns, which fall under the heading of what Harsanyi calls "psychological difficulties",[48] can be effectively eliminated. If for some reason an individual is quicker than others to express her emotions, we may suppose that we can have independent evidence for this, and can compensate in our judgments about her welfare.

There may also be concerns about how much of an individual's expressive behaviour is due to her satisfaction with her present situation, and how much is due to expectations regarding which of her preferences will be satisfied in the future. If our goal is to estimate the extent of currently satisfied preferences, then we need some way of isolating and excluding that component of individuals' behaviour resulting from expectations of future gain; whereas if our goal is to estimate degree of preference for future states of affairs, we need to discount behaviour due to present satisfaction. Perhaps

---

[47]For some doubts about the legitimacy of this charge see Ilmar Waldner, "The Empirical Meaningfulness of Interpersonal Utility Comparisons", Journal of Philosophy Vol. LXIX, No. 4 (1972), pp. 87-103.

[48]Ibid., p. 16.

this difficulty can be met simply by taking the states or outcomes which expressive behaviour is supposed to indicate degree of preference for to be temporally extended. In the limiting case, the outcomes might be full possible worlds, with histories extending into the past and future, and individuals' expressive behaviour would be understood as indicating overall degree of satisfaction with the world in which they are situated (or perhaps as indicating the weighted extent of satisfaction with the actual and other possible worlds, if there is some uncertainty as to what the actual past and future is/will be). In any case, let us grant that the difficulty can be managed.

Still, and notwithstanding the impressive array of charitable assumptions we have so far made on behalf of the advocate of behavioural comparisons, the proposal before us suffers from a critical defect. Suppose that we judge someone to be more miserable than one of her fellows in accordance with the criteria Little sketches. We are to conclude that the cumulative strength of her satisfied preferences is somewhat less than the strength of his satisfied preferences, and thence by consulting their respective interval charts of preference intensity, get a rough idea of how the strength of any of her preferences stacks up against any of his.

But now consider: Is the individual in question miserable because the intensity of her currently satisfied

preferences is relatively low? Or because the intensity of her unsatisfied preferences is relatively high? Isn't it possible, e.g., that the cumulative strength of her currently satisfied preferences is greater than the strength of all of his preferences put together, satisfied as well as not; but that she is miserable because of the still greater intensity of her frustrated preferences? The evidence seems consistent with both the original hypothesis and this one, even though they yield radically different schedules of comparison.

Of course, if we already knew what the intensity of her unsatisfied preferences was (the "absolute" intensity, that is, rather than the intensity relative to her satisfied preferences), then we would have no trouble deciding what the strength of her satisfied preferences was, and thus would have a ready answer to the question of which of the two hypotheses is correct. But if we had that information we wouldn't need to bother with behavioural evidence; we would already have discovered an objective unit of comparison, and could simply read the intensities of people's preferences right off their utility functions. Without making assumptions about the intensities of their unsatisfied preferences, the behavioural evidence will not so far as I can see decide between the competing hypoth-

eses.[49]

Harsanyi seems to recognize this difficulty, but evidently supposes that it can be met by invoking a "principle of unwarranted differentiation":

> If two objects or human beings show similar behaviour in all their relevant aspects open to observation, the assumption of some unobservable hidden difference between them must be regarded as a completely gratuitous hypothesis, and one contrary to sound scientific method....Thus in the case of persons with similar preferences and expressive reactions we are fully entitled to assume that they derive the same utilities from similar situations.[50]

This will not do, however. We need to distinguish between the broadly physicalist stance which is arguably a precondition of sound science, and the overt behaviourism which Harsanyi would have us adopt as a corollary. Let us grant that exactly similar individuals -- i.e., atom-for-atom replicas, in exactly similar circumstances -- have exactly similar preferences and preference strengths. This much may plausibly be taken to be an expression of sound

---

[49]Ultimately I shall be claiming that the two hypotheses are simply grammatical variants expressing the same underlying facts -- that there is no difference between someone being miserable in virtue of the low intensity of her satisfied preferences and in virtue of the high intensity of her unsatisfied preferences. That answer is hardly available to the proponent of behavioural comparisons, however, since it requires an extended argument in support of the claim that preference intensities are irreducibly relational in nature; from which it follows that interval scales provide all of the information about individuals' preference strengths which it is possible to have, and hence that behavioural data are irrelevant to judgments of interpersonal welfare once we have personal interval scales in hand.

[50]Ibid., pp. 15-16.

scientific method, since denying it would apparently amount to denying that preferences and/or preference intensities are at base physical entities/characteristics, and hence fit subjects for study by science as we know it.

In contrast with this minimal physicalist assumption, it should be abundantly clear that Harsanyi is advancing a substantive thesis about the functioning of individuals' psychologies. Given that the "similar preferences" to which he refers in the last sentence of the above-quoted passage are by hypothesis similar only as measured by the vN.M. procedure, his invocation of the principle of unwarranted differentiation amounts to the claim that if two individuals have identical vN.M. interval utility functions, a difference in the absolute strengths of their preferences would necessarily result in a difference of expressive behaviour in some circumstances. (In Jevons' terms: It is a presupposition of sound science that two individuals' preferences cannot differ "in a like ratio in all directions" without there being some behavioural upshot.) Thus if an individual is miserable in virtue of the low intensity of her satisfied preferences, rather than because of the high intensity of her unsatisfied preferences, then there must according to Harsanyi be something about her behaviour which reveals this.

This substantive claim is surely not defensible on mere grounds of scientific methodology. To reiterate: On

the assumption that preferences have absolute magnitudes at all, as physicalists we are committed to holding that the magnitudes of two individuals' preferences may differ only when there is some physical difference between them. But not necessarily a behavioural difference. The latter conclusion would of course follow if one were in addition a logical behaviourist; but being one of those is hardly a precondition for engaging in scientific inquiry.

There is a further point which needs emphasizing here. Quite apart from his attempt to masquerade behaviourism under the banner of science, Harsanyi's strategy for countering scepticism with respect to interpersonal comparisons is not very helpful. To see this, note that his psychological thesis might nonetheless turn out to be true. I.e., it might turn out that, as a matter of contingent psychological fact, individuals with similar interval utility functions never differ in the absolute strengths of their preferences without displaying some behavioural differences; this possibility can obviously not be ruled out a priori, any more than it can be confirmed in the manner Harsanyi suggests.

The important thing to notice, however, is that regardless of whether the possibility should turn out to be actual, the kinds of behavioural facts that Little and Harsanyi believe we do rely on in making interpersonal comparisons do not yield an answer to the question of

whether someone is happy or miserable in virtue of the weight of her satisfied preferences, or in virtue of the weight of her unsatisfied ones. Thus even if Harsanyi is right in thinking that there must be some behavioural facts which distinguish the two hypotheses, we do not appear to base our judgments of interpersonal welfare on them. To the extent that our ordinary judgments are understood as being about absolute strengths of preference, then, we can only conclude with Robbins that these judgments <u>are</u> subjective: they rest on assumptions, unsupported by observations of behaviour, about the absolute weight of individuals' unsatisfied preferences. And since what little evidence we have for believing in the possibility of objective interpersonal comparisons stems from our as yet unsubstantiated confidence in our ordinary judgments, a demonstration that such judgments are in fact irredeemably subjective is as good as a demonstration that we have no reason to believe in the possibility of objective interpersonal comparisons at all -- hence no reason to think that Harsanyi's psychological thesis is in fact true. Unless the proponents of behavioural comparisons are able to produce at least a sketch of the more fine-grained distinctions of behaviour that would be required to discharge assumptions about the strengths of individuals' unsatisfied preferences -- or at any rate, some argument to suppo , the claim that these finer distinctions do exist -- it would

seem that we are fully entitled to remain sceptical about the possibility of comparing welfares on the basis of behaviour.

5. Direct Physiological Comparisons

If behaviour does not seem capable of providing us with an interpersonally significant measure of preference intensity, then why not go straight to the source, and compare the physiological bases of individuals' preferences? This is roughly the suggestion that Richard Brandt adopts in arguing the case for interpersonal comparisons in A Theory of the Good and the Right.[51] Actually, Brandt rejects preference-based accounts of welfare in favor of a pleasure-based account, and so his proposal is concerned in detail with comparing the "net enjoyment" or cumulative "pleasantness and unpleasantness" of individuals' experiences. We have already discovered reasons for thinking that preference-based accounts of welfare are better suited to utilitarian purposes than pleasure-based ones, but before turning to Brandt's defense of direct physiological comparisons, it may be instructive to examine his reasons for eschewing the former.

The difficulty which Brandt locates with preference-based accounts is perhaps best brought out by contrasting them with his own view, which he refers to as the "hap-

---

[51] Oxford: Clarendon Press, 1979, Ch. XIII.

piness theory". The happiness theory identifies an individual's welfare over time with the pleasantness of her experiences, where the degree of pleasantness of an experience is determined by the extent to which it makes the agent want to continue it (or repeat it) for itself. Thus Brandt's theory does ultimately make reference to strengths of desire in determining welfare. The happiness theory is importantly different from a preference- or desire-based theory, however, in that it is not the satisfaction of desires as such which counts towards an individual's welfare. Rather, what counts are moments of happiness or "positively valenced" (i.e., desired) experience. Crucially, it is not the fact of their being positively valenced which makes such moments of experience count on Brandt's view. The desires which are relevant are limited to those desires for the continuance of an experience which are caused by the experience at the time, and their relevance is limited to determining the magnitude of pleasantness or unpleasantness of the experience.[52]

The ingenuity of Brandt's happiness theory cannot be doubted. One of its virtues is that it deftly avoids a problem which has sometimes been thought to afflict pleasure-based accounts generally: namely, the problem of estimating the extent to which pleasures which differ

---

[52]See ibid., Ch. II; and "Two Concepts of Utility" in Miller and Williams, op. cit., pp. 169-85, esp. Section 6.

greatly in quality contribute to someone's welfare. How may we gauge, e.g., the degree of pleasantness of sensual pleasures like those of eating ice cream or sipping a cool drink on a hot afternoon, in comparison say with intellectual pleasures like those accompanying the study of philosophy? Brandt's answer is simple: the pleasurableness of an experience is determined by the strength of the accompanying desire which one has for the experience to continue. Since strengths of desires for the continuance of experiences are presumably commensurable in a way in which the experiences themselves are not, such desires may serve as the common coin by which the relative contributions of different kinds of experience can be estimated.[53]

Brandt's happiness theory enjoys an important advantage, then, over other pleasure-based accounts of welfare. Nevertheless, given that it is <u>not</u> the fact of their being

---

[53]It might be wondered, though, whether estimating the contribution of different kinds of experience in this fashion provides the correct answer. Most of us are familiar with experiences which are highly positively valenced at the time, even though in calmer moments we might prefer not to undergo or have undergone them. Inebriation is often such an experience. Nor need one's reflective preference for not being inebriated too often depend on the unpleasant effects which sometimes follow. Even without the subsequent instrumental disutility (tequila works well for me), one might reflectively prefer to engage in other activities (for their own sake, let us add); and one might well prefer this even though those other activities would be less positively valenced at the time. To the extent that this is true the example serves as a nice illustration of how pleasure and preference satisfaction may part company in real-world as opposed to merely theoretical settings.

desired which makes moments of pleasurable experience count towards a person's welfare, it should be clear that the happiness theory is a version of what we earlier referred to as hedonistic intuitionism, and as such the utilitarian can buy into it only at the cost of giving up the attempt to thoroughly ground moral value. Moreover, although the happiness theory does neatly resolve the problem of incommensurable experiences, it suffers from other defects in common with all pleasure-based accounts. Like all such accounts, Brandt's theory defines welfare in terms of purely subjective aspects of individuals' experiences, leaving out the objective or mind-independent component which is a characteristic concern of preference-based theories. The happiness theory implies that we can increase someone's welfare merely by inducing positively valenced mental states, thereby raising the spectre of making people "better off" by forcibly wiring them up to various sorts of pleasure contraptions -- a possibility which strikes many people as repugnant from both the moral and personal points of view.[54]

Still, however suspect pleasure-based accounts of welfare must remain from a utilitarian perspective, we have not yet addressed Brandt's reasons for thinking that preference-based accounts are faultier still. The chief

---

[54]See e.g. Nozick's discussion of the "experience machine" in Anarchy, State, and Utopia (New York: Basic Books, Inc., 1974), pp. 42-5.

difficulty with the preference-based view, according to Brandt, is that there is no plausible way of formulating a coherent program for promoting the satisfaction of an individual's preferences or desires. Suppose that we wish to determine which of two actions available to us, a and b, will best promote A's welfare. On the happiness theory, what is required in order to arrive at this determination is simple enough in theory, however complicated it might be in practice. Assuming that we can make reliable predictions concerning A's experiences at each moment of her life, we can graph the relative enjoyments produced by a and b along a temporal axis, with points above the axis representing moments when her experiences consequent on doing a would be more positively valenced (less negatively valenced) than would her experiences consequent on doing b (the height above the axis corresponding to the difference in valence of the hypothetical experiences), and points below the axis representing moments when her experiences consequent on b would be more valenced. We can then simply calculate the area under the curves, with action a contributing more to A's happiness or net enjoyment over her lifetime than action b just in case the total area above the line is greater than the area below it.[55]

---

[55]See A Theory of the Good and the Right, p. 248, or "Two Concepts of Utility", pp. 175-6. This procedure obviously depends on our ability to perform intrapersonal comparisons, comparisons of the strength of A's desires at different times, on pain of being unable to relate the

In contrast with the theoretical simplicity of this sketch, how might we go about promoting the satisfaction of an individual's desires? In Brandt's view the problem here is two-fold. In the first place, many of the desires which an individual has at a time t are for states or events which will occur (or which have occurred) at some time other than t. Secondly, an individual's desires, and in less extreme cases their intensities, are subject to change over time.

Given these facts, there does not appear to be any plausible way of focusing on A's preferences at some particular moment in order to determine how best to promote her welfare. Suppose e.g. that action a would result in some event e occurring at t' in the future, while b would result in a different event f occurring at that time. Imagine further that A currently prefers that e occur

---

distance of the curve above or below the line at one time to the distance at other times. In my view the "problem of intrapersonal comparisons", if we may call it that, is structurally and substantively identical to the interpersonal version. Brandt discusses intrapersonal comparisons on pp. 253-7 of A Theory of the Good and the Right; without rehearsing the details, I think that the objections to be raised below against his proposal for performing direct interpersonal comparisons could be aimed with as good effect at his account of intracomparisons -- preference intensities simply cannot be compared directly, for reasons which will emerge in the next chapter, whether they attach to the preferences of one individual at different times, or to those of different individuals at the same or different times. For a brief discussion of the relation between intra- and interpersonal comparisons see Roy A. Sorensen, "Did the Intensity of My Preferences Double Last Night?", Philosophy of Science 53 (1986), pp. 282-5.

rather than f at t′; but that in the interim this preference will reverse itself (regardless of which action is performed), so that when t′ finally arrives A will prefer f to e. Ignoring for the moment other consequences of the two actions, what should we conclude about the relative efficacy of a and b in promoting the satisfaction of A's desires? Giving priority to her present desires would of course lead us to favor a, whereas focusing on the desires she will have at the time of the hypothetical events would give the nod to b. But why should we do either? If it is the satisfaction of A's desires as such that we are aiming at, it is difficult to see why her present desires should be granted priority over her future ones, or vice versa.

An alternative might be to take into account the desires which A has at every moment of her life. Assuming that we can know for each moment what her preferences are (have been/will be) regarding the occurrence of e and f at t′, we could graph those preferences along a temporal axis in a manner similar to that suggested by Brandt for the corresponding happiness calculation. Points above the line would correspond to moments when A prefers e's occurrence at t′ to f's, with height above the line indicating degree of preference, and similarly below the line for moments when she prefers f's occurrence to e's. Action a would then contribute more to A's welfare than b would just in case the area demarcated by portions of the curve above the

line is greater than the area below it.

The difficulty with this suggestion is that it appears to grant undue weight to A's past (and perhaps distant future) desires. Past desires pose no difficulty for the happiness theory because the only desires relevant to the calculation are those accompanying the experiences that one is contemplating bringing about at a given time. Desires in general are typically for states or events occurring at other times, however, and thus it is quite possible to satisfy an individual's past desires, whether or not she currently possesses them. If it is the satisfaction of desires as such which counts towards A's welfare, then presumably all of her desires should be included in the calculation. But do we really want to be bound to acting on A's past desires, even ones which she has long since ceased to possess? (Suppose that for a long time A preferred e's occurring at t' to f's -- long enough for this past preference to outweigh her present and future ones in accordance with the calculation sketched above -- but that she currently and for the remainder of her life will prefer f to e. Do we really think that this should prompt us to favor a rather than b so far as A's welfare is concerned? That would appear to be a perfect recipe for making her miserable for the rest of her life.)

Worries such as these have led Brandt to doubt the intelligibility of preference-based accounts of welfare. I

suggest that the worries are misplaced. They result from implicitly attempting to re-interpret the preference-based view in intuitionistic terms.

Let us set aside desire satisfaction for the moment, and focus on the central features of the utilitarian account. The principle of utility advises us to rank actions, policies, etc. in accordance with the extent to which their effects coincide with what people value. Now, people's values do certainly change over time. But this is perfectly consistent with the principle of utility, as I understand it. Indeed, it seems to me to be a straightforward consequence of the principle that, to the extent that individual values are subject to change, so too are utilitarian evaluations. (We must resist the temptation to detach moral value from individual values in a way which would allow the former to attach to actions or other events or states of affairs once and for all. From the utilitarian perspective value is inevitably a relation between valuer and thing valued. It thus makes no sense to attribute value of any sort to a thing, independently of specifying a person or persons for whom it has value at some particular time.)

Thus the crucial question so far as application of the principle of utility is concerned is simply whether it is possible to identify, at the moment of ranking, what things individuals value. But this is trivial: The things which

an individual values at a particular time are precisely specified by the preferences she has at that time[56] -- preferences regarding how her life will go in the future, and more generally preferences regarding how the world will go, since the things which people value are typically not restricted to aspects of their own lives. (People also have preferences regarding things in the past, of course, but such preferences will not affect the utilitarian's ranking, since it is a "consequence" of whatever action one might now perform that the past will be just what it was.) An individual may have had different preferences in the past, and she may come to hold different ones in the future, both regarding how the world will go and how it has gone; but such past and future preferences concern what things she did or will value, not what things she does.

In ranking actions, then, the utilitarian will focus on individuals' present preferences regarding how the world will go. She will do so in the knowledge that preferences are subject to change over time -- perhaps even change as a result of the actions she is contemplating -- and that such changes may dictate future revisions in the evaluation of present policy. There is nothing especially puzzling about this. A single individual may realize, in full knowledge of her situation, that her preferences will change at some

---

[56]More precisely, by her reflective and informed preferences at the time; see n.60 below.

point in the future (indeed, who doesn't realize this to some extent?), and may favor a world with revised preferences to worlds in which her preferences remain fixed. To take the most extreme case, she may even rationally undertake to revise certain of her preferences, realizing that if she is successful she will come to assess her present values and actions very differently than she now does.[57] There is nothing essentially irrational in the idea that individuals may presently value, among other things, coming to value things which they do not currently value, or ceasing to value things which they currently do. To the extent that someone's present preferences are in fact structured in this way, neither is there anything unintelligible in the utilitarian's rankings of policy

---

[57]For an interesting discussion of when it might be rational to undertake to revise one's preferences, see Duncan MacIntosh, "Retaliation Rationalized", ms. presented to the annual meetings of the Canadian Philosophical Association, Windsor, May 1988.

It might be wondered whether the possibility of possessing considered preferences regarding one's future preferences could lead to a kind of circularity or instability. Could one rationally prefer now to acquire certain preferences in the future, the satisfaction of which would in turn involve reacquiring the old preferences? We can probably make reasonable sense of the suggestion insofar as the original preference is understood as being instrumental. I leave it to the reader to construct a plausible story, but peculiarities aside, I do not think that the case poses any special difficulties for the preference-based view. Insofar as the original preference is understood as being intrinsic rather than instrumental, it's not clear to me that the possibility is a real one, rational or otherwise. It amounts to the suggestion that one might value coming to be the sort of person who values being the sort of person that one now values being. What sort of person is that? (Perhaps every sort.)

following suit. (So far as A is concerned: If she is fully cognizant of the fact that her present preference regarding e and f will reverse itself prior to t', and for whatever reason she nonetheless persists in ranking a above b, then the utilitarian cannot in consistency do otherwise.)

Brandt briefly considers this proposal -- that on the preference-based view, what one does is evaluate actions at any time in accordance with the preferences which individuals possess at that time -- but rejects it as being "arbitrary and unsatisfactory compared with the original tidy goal of satisfying a person's desires, past and future, maximally, based on a picture of all desires he will have at every moment of his life."[58] Unfortunately he does not go on to explain why he considers the program in question to be arbitrary and unsatisfactory in comparison with attempting to maximally satisfy all of a person's past, present, and future desires. Obviously the judgment does not turn on ease of application, since in this respect the latter policy is far less tidy than the former, given that its implementation would require us to ascertain not only individuals' present preferences, but their preferences at every past and future time as well.

It is difficult to be sure what Brandt's precise reasons are for thinking that the policy of focusing on individuals' present preferences is unsatisfactory, but I

---

[58] A Theory of the Good and the Right, p. 251.

suspect that they ultimately derive from an underlying intuitionism regarding questions of moral value. What separates the utilitarian approach from others is its clear refusal to countenance moral value of any sort which is not directly traceable to what particular individuals actually value. It seems to be a basic presupposition of Brandt's reading of the preference-based view, on the other hand, that desire satisfaction as such is morally valuable. On this interpretation, what the morally conscientious person wants ideally to do is to bring it about that there is as much desire satisfaction in the world as possible over time. The fact that individuals' desires are subject to change over time threatens the intelligibility of this program, however: there appears to be no satisfactory way of calculating how much desire satisfaction exists at a given moment, since the amount present varies according as we focus on people's preferences at different times. Attempting to circumvent the difficulty by taking into account the preferences which individuals have at every moment of their lives will yield an unequivocal answer concerning how much desire satisfaction exists during any interval, but it also yields an intuitively unsound guide to policy, one which would grant undue weight to individuals' past preferences. Neither can the problem be avoided by focusing on the preferences which people have at the time of appraisal; this proposal just misses the point,

since it provides an answer to the different question of how to maximize the satisfaction of individuals' present desires, not the question of how to maximize desire satisfaction per se, with all past, present, and future desires taken into account.

From a utilitarian perspective, this construal of the preference-based view is fundamentally flawed. The problem with it is that rather than taking advantage of the opportunity to ground moral value in individual values, it in effect postulates the existence of a kind of stuff, desire-satisfaction, which is assumed to have intrinsic moral worth. The point cannot be emphasized too strongly -- nothing has value of any sort for the utilitarian, except insofar as someone values it at some time. The utilitarian does indeed hold that states of affairs are morally valuable just to the extent that they are valued, and hence that bringing about such states -- satisfying desires "as such" (i.e., satisfying someone's present desires regardless of whether this will produce pleasure or anything else which at one time or another has been regarded as morally valuable) -- is a (rather, the) morally worthwhile activity. But satisfying desires as such in the sense of attempting to maximally satisfy desires which individuals no longer possess is from the utilitarian perspective an essentially worthless activity (or worse!). I conclude that Brandt has failed to produce reasons for

thinking that preference-based accounts of welfare are inadequate, at any rate for purposes of foundational ethical theorizing.[59] [60]

We come finally to Brandt's proposal for performing

---

[59] Allan Gibbard has recently echoed some of Brandt's concerns regarding preference-based accounts of welfare in "Interpersonal Comparisons: Preference, Good, and the Intrinsic Reward of Life" (Jon Elster and Aanund Hylland, eds., Foundations of Social Choice Theory, Cambridge University Press, 1986, pp. 165-93). He asks us to imagine someone who initially prefers a life of austere religious contemplation, but mistakenly thinks that commitment to such a life will be fostered by experience in a worldly university:

> If as a result of his university experience he comes to lead a life he values as he leads it, but would have despised when he initially chose the university, can we conclude that life has been bad for him? If not, his initial...preferences do not measure his good or his welfare, and so cannot reasonably be taken as his 'utility' for ethical purposes.

I agree that we would not be inclined to think, after the fact, that his life had gone badly. Indeed, this is just what the preference-based view predicts. After the fact the person values his life as he leads it (and presumably, as he has led it), and the utilitarian takes such judgments at face value. This could hardly justify taking steps beforehand to guide the person into a life which he in fact despised at the time, however; that would amount to a recipe for justifying all sorts of brainwashing techniques.

[60] There are a number of other issues relating to preference-based accounts of welfare and the foundations of utilitarianism which cannot be given fully adequate treatment here. Perhaps the following brief remarks will suffice:

1. Many theorists distinguish individuals' explicit or "revealed" preferences from their reflective and informed preferences, and hold that it is the latter which are relevant to ethical theorizing. (See e.g. John Harsanyi, "Rule-utilitarianism and Decision Theory", Erkenntnis 11 (1977), pp. 25-53.) The point is well taken. In Gibbard's example, the person who values a life of religious contemplation may in a sense be said to prefer experience at a worldly university over other alternatives (that is what he says he prefers, that is what he chooses). But this

direct interpersonal comparisons. Having rejected his reasons for adopting a pleasure-based account of welfare, we shall nevertheless follow his lead in discussing the comparison of individuals' pleasures or enjoyments rather

---

revealed preference is misinformed, based on a mistaken view of the effects of the experience. The preferences with which we are concerned are presumed to exist prior to, and, in conjunction with beliefs, to guide agents' behaviour. Thus his true preference, the one relevant to the utilitarian's ranking of actions at the time, is something different.

2. Authors in addition to Brandt and Gibbard have suggested that individuals' welfares cannot be unqualifiedly defined in terms of their preferences, including their preferences regarding others' welfares, since that would make it trivially true that people always act in their own interest -- it would be definitionally impossible for someone to sacrifice her own welfare for the sake of others. (See e.g. Mark C. Overvold, "Self-Interest and the Concept of Self-Sacrifice", Canadian Journal of Philosophy 10 (1980), pp. 105-18; and "Self Interest and Getting What You Want" in Miller and Williams, op. cit., pp. 186-94.) The worry rests on a misapprehension of the structure of utilitarian theory. Note first of all that it is definitionally true, barring mistakes of one sort or another, that individuals act to bring about what they value. That being said, there is no bar to the utilitarian distinguishing between self-regarding and other-regarding values; people who possess and act solely on self-regarding preferences are selfish people, plain and simple, whereas those who act on other-regarding preferences may in a perfectly good sense be said to be "setting others' interests before their own", and are likely to be in for a good deal of praise on any utilitarian account. The distinction between self- and other-regarding values is one which should be drawn at the level of the utilitarian's theory of obligation, however, not located within the foundations of the theory (see Ch. 5 below). The utilitarian identifies moral value with whatever it is that individuals value (she refers to this as their "welfare", though the term is dispensable if linguistic intuitions balk too loudly), whether this happens to be aspects of their own lives, or of others', or of no-one's at all. To rank states and actions on any other basis is inevitably to impose an intuitionistic standard to some extent.

3. Commentators have sometimes wondered whether utilitarianism should be given a "classical" or an "average"

than the strengths of their preferences per se. Brandt notes that his proposal could be adapted without much difficulty to a preference-based view of welfare, since on the happiness theory comparing individuals' pleasures

---

formulation; i.e., whether the theory is concerned with promoting the sum total of individuals' welfares, or with promoting their average welfare. (The two are of course equivalent given a fixed population.) In view of our discussion of the intuitionistic basis of Brandt's worries about preference-based theories, it should be clear that the averaging formulation is the correct one. To think otherwise is implicitly to suppose that welfare as such is morally valuable, that the utilitarian's fundamental concern is to get as much of it into the world as possible, and that since people happen to be the locus of the stuff, actualizing as many possible individuals as one can (so long as this does not result in too steep a degradation of present individuals' welfares) is a good means of getting more of it into the world. This is a mistake. The welfare of merely possible individuals is as intrinsically worthless as everything else. (Cf. Smart, "An Outline of a System of Utilitarian Ethics", pp. 27-8. Smart suggests that the choice between the total sum and average formulations revolves on nothing more than the utilitarian's intuitions. He is however a virtual paradigm of the hedonistic intuitionist as we characterized that position above, and hence it should come as no surprise that his own intuitions favor the total sum view. In my view, hedonistic intuitionists should not be counted as utilitarians at all. If this seems like a perverse suggestion, recall that G.E. Moore also perversely dubbed his own brand of intuitionism "utilitarian".)

4. If there is a puzzle regarding preference-based accounts of welfare, it has to do not as Brandt and Gibbard suggest with formulating a coherent program for promoting the satisfaction of a single individual's preferences, but rather with reconciling potential conflicts between the preferences of present individuals and those of future ones (not merely possible individuals, mind -- they don't count for anything -- but actual future ones). If the utilitarian were to rank actions in accordance with only the present preferences of extant individuals, then the values of future persons would count for nothing in moral assessment inasmuch as they diverged from the values of those present. The usual solution to the "problem of future generations" -- that the concerns of future individuals will be adequately looked after in present moral rankings

reduces to comparing the strengths of their desires for the continuance of experiences. The proposal strikes me as good deal more intuitively plausible when presented in terms of comparing the pleasantness and unpleasantness of experiences, however; if anything, the difficulties that I want to raise for it would arise more starkly for a version couched directly in terms of preferences. Hence for the sake of giving the proposal as sympathetic a hearing as possible, we will examine it simply as presented.

Brandt's idea is that, at least for experiences which are fairly closely tied to changes in an individual's physiological condition, we can roughly gauge the intensity of the experience in question by measuring changes in its physiological correlate. Thirst, e.g., is triggered by chemical changes in the blood stream due to dehydration, and roughly increases with dehydration. Thus we can test for degrees of thirst by measuring levels of dehydration. Moreover, under certain conditions (of which, more in a

---

via extant individuals' preferences regarding the welfare of their offspring, and their offspring's offspring, etc. -- does not address the nagging worry that future persons' values do count for something in present assessments, particularly when they happen not to coincide with the concerns of present individuals. I think that the answer here must be that the utilitarian does not perform rankings solely on the basis of the present preferences of present individuals; she also takes into account something like the considered preferences that future individuals would have were they now present. But that is a difficult counterfactual to evaluate, and I propose to keep the can opener as far away from this particular tin of worms as possible.

moment) we may take this sort of test to have interpersonal significance with respect to degrees of pleasantness or unpleasantness. Of course, tests of this nature are plausible if at all only for experiences which are closely connected with known physiological responses. But on the assumption that the pleasantness/unpleasantness of all of an individual's experiences may be ranked on personal interval scales (Brandt's version of our assumption that we may in principle have access to exhaustive interval scales of preference for each individual), an objective comparison of levels of dehydration may fund us with full interpersonal comparability.

Since the goal is to arrive at an interpersonally significant measure of the pleasantness of experiences, we must be careful to control any features of people whose levels of dehydration are to be compared which might make us suspect that the unpleasantness of their thirst given a certain level of dehydration is augmented or diminished. Brandt mentions a number of conditions that individuals must be matched for, including recent experience and habituation, conditioned anxiety responses, pain thresholds, defective thirst-triggering systems, and bodily weight.[61]

Attempting to match people for some of these features would appear to be question-begging (on what grounds might we hold that one person's pain threshold is roughly the

---

[61] Ibid., p. 262.

same as another's, e.g., if not the supposition that the unpleasantness of certain physical injuries is roughly equivalent for them?), but we shall let the difficulty slide. We do however need to extend Brandt's list of conditions somewhat. For one thing, the pleasantness or unpleasantness of experiences may be considerably affected by such things as whether one is distracted or preoccupied with other things, so we need to somehow insure that individuals whose levels of dehydration are to be compared are also matched in this regard (though how we might go about controlling for level of attention to an experience I have no idea).

Much more importantly, there is reason to think that we cannot take levels of dehydration to be a good interpersonal indicator of the unpleasantness of thirsts unless individuals are matched as well for their desires concerning higher-level experiences, ones which are less closely allied with simple physiological responses. Notice that at the level of common sense, a moderate degree of thirst may not be at all unpleasant when one may anticipate consuming a cold beer in the near future. (But this assumes that one actually <u>likes</u> cold beer, and will derive pleasure from consuming it; otherwise the unpleasantness associated with one's thirst is likely to be exacerbated by the prospect of having nothing to drink afterwards but such a vile liquid.) In terms of the details of Brandt's motivational account of

pleasure, the extent to which one is motivated to discontinue an experience such as thirst depends not only on the experience itself, but also on one's perception of the alternatives to its continuance, and on one's preferences regarding the same (which may in turn depend on a host of other factors; e.g., if I believe that drinking beer reduces one's life span, this is likely to have a bearing on the extent to which the prospect of consuming a cold beer influences the unpleasantness of my present thirst -- assuming, of course, that I derive pleasure from the prospect of increased longevity).

According to the happiness theory the pleasurableness of an experience is determined by the extent to which it causes one to desire its continuance at the time, and what I am in effect pointing out is that such things as one's beliefs, expectations, and higher-level likes and dislikes concerning alternatives can be expected to have an impact on the causal processes at issue, and hence on the strength of the resulting desire for continuance. It may be objected that Brandt has a ready response to this claim, however, since on the happiness theory the pleasantness of experiences is supposed to be determined by how they influence the valence of their continuance <u>for themselves</u>. Brandt introduces this qualification in order to preclude a kind of "double-counting" or "over-counting" in estimating the pleasantness of experiences, given that one may desire

an experience both for itself and for certain instrumental effects. It might be supposed that the qualification can be put directly to good use in the present context, however; for while it is no doubt true that the extent of my motivation to discontinue an experience such as thirst depends to some extent on my preferences regarding perceived alternatives, it may not seem implausible to hold that such preferences do not causally impact on how strongly I am motivated to discontinue the present experience of thirst "for itself".

I don't think this will help, for the simple reason that I don't think it's possible to make clear sense of the notion of "desiring the continuation of an experience for itself" in complete abstraction from any and all alternatives. Brandt doesn't explicate the notion at any length and I'm unsure how he would proceed, but consider the following dilemma: Suppose that at some time the continuation of an experience E is positively valenced for me, and then I suddenly become aware of what I regard as a preferable alternative. In this case E will shift from being positively to negatively valenced and I will actively seek to discontinue it, with the degree of negative valence dependent on how strongly I prefer the alternative.[62] Now,

---

[62]Note that the shift in valence here is, or at least may be, independent of any supposed instrumental effects of E and its alternative. Suppose e.g. that I am contentedly quaffing a run-of-the-mill domestic beer, and then suddenly become aware that the host has a good stock of imported

has the valence of E's continuance "for itself" similarly shifted from being positive to negative? If so, then my point is demonstrated; the very direction of valence of an experience's continuance for itself, much less the degree, depends on one's preferences regarding the perceived alternatives, and hence we cannot legitimately infer similar degrees of valence for the continuance of physiologically similar experiences without first matching individuals for their preferences and preference strengths regarding alternatives.

If not, then the notion of "degree of valence of an experience for itself" must evidently be understood counterfactually, as referring not to the actual valence of E's continuance, but rather to the valence that E's continuance would have under certain circumstances. But what circumstances? The problem here may be formulated in terms of a further dilemma: Either the counterfactual circumstances which are relevant to determining the valence of E's continuance for itself are ones in which there are no perceived alternatives to E, or ones in which there are. If the former, then every experience will turn out to be positively valenced for itself (or perhaps neutrally valenced), since in the absence of any perceived alternatives I won't display a tendency to discontinue any experience. If the counterfactual circumstances do include

---

ales behind the bar.

perceived alternatives, on the other hand, we're back to the original problem: whether I'm disposed to continue or discontinue E and to what extent depends on just what the alternatives are, and on how favorably I view them. Place a glass of salt water in front of me and I won't be inclined to assuage even an intense thirst. But set up a glass of cold beer....[63]

There is a moral to be drawn here concerning the nature of value. We noticed above that from the utilitarian perspective, value is always a relation between a valuer and the thing valued. The discussion just concluded suggests that this proposition should be modified somewhat, however. Fully cashed out, value is not a two-part

---

[63]Brandt actually defends a dispositional account of wants/valences, the centrally relevant counterfactuals being ones which concern a person's "tendency" to perform some action if it were to occur to her that doing so would bring about the desired outcome; see ibid., pp. 25f. He resists a probabilistic interpretation of the tendencies in question, on grounds that we may want to say that someone has a strong tendency to perform some action, even though she is unlikely to perform it because of the presence of a stronger contrary tendency; rather, tendencies are to be understood by reference to the role they play in psychological laws (e.g. "An agent will actually perform an action A if and only if the tendency to perform A is stronger than the tendency to perform any other action B"). Thus I presume that in the present example he would want to say that E remains positively valenced for me "in itself", though its alternative is more strongly valenced. His invocation of psychological laws appears, however, to acknowledge the very point I am stressing; for such laws are either identical to or intimately connected with counterfactuals relating the likelihood of an agent's performing some action to the strength of her tendency to perform that action in relation to the strengths of her tendencies to perform alternative actions.

relation between valuer and thing valued; it is rather a three-part relation between the valuer and two objects or events or states of affairs. Properly speaking I do not value objects or states or experiences all on their own; I rather value them (and disvalue them) in relation to other objects and states and experiences. That is one reason why discussions of value are ultimately best carried on in terms of individuals' preferences; the latter make explicit the three-part nature of the value relation, whereas speaking of individuals' wants or desires or "valences" for outcomes has a tendency to obscure this fact.

Morals aside, it is apparent that despite a certain initial plausibility, Brandt's proposal for performing direct physiological comparisons will not bear close scrutiny. The more pleasure I take in the prospect of consuming a beer, the less unpleasant a moderate degree of thirst, and similarly for you. Thus in order to arrive at an objective comparison of the unpleasantness associated with our thirsts, we require some independent means of determining the relative pleasures we derive from the prospect of consumption. But we don't have an independent means of comparing the relevant pleasures; to suppose that we did would be to suppose that we had already solved the units of comparison problem, and hence that there was no need to bother with measuring dehydration levels in the first place. Physiology no more than behaviour, it seems,

will provide a satisfactory basis for interpersonal comparisons.[64]

---

[64] In discussing the "serious logical justification" for his proposal (ibid., p. 262) Brandt avers the dual supposition that the unpleasantness of individuals' experiences is caused somehow, presumably by events in their brains; and that like causes have like effects, so that if the complex cause of the unpleasantness of one person's experience is duplicated in another person, we may assume that the unpleasantness of the second person's experience is the same.

This much we may certainly grant (we already granted it above, in our discussion of Harsanyi's proposal): physically indistinguishable individuals may be presumed to have exactly similar preferences and preference strengths (whatever that finally comes to), including preferences for the continuation of experiences. This basic physicalist presupposition will no more substantiate direct physiological comparisons than it did behavioural ones, however. If we could identify people who were physical replicas, we wouldn't need to bother measuring levels of dehydration to know that whatever the degree of thirst of the first and however unpleasant, the second would be just as thirsty and receive just as much displeasure. The real problem of interpersonal comparisons would remain: For individuals who are obviously not physical replicas, and who may differ in their preferences regarding alternatives to the continuation of various experiences, inferring similar unpleasantness from similar levels of dehydration would beg the question by assuming that their preferences regarding the alternatives were the same.

CHAPTER 3: INDIRECT INTERPERSONAL COMPARISONS

> "There once lived on the banks of the Indus River an ancient Persian by the name of Al Hafed. He owned a lovely cottage on a magnificent hill, from which he could look down upon the glittering river and the glorious sea; he had wealth in abundance, fields, grain, orchards, money at interest, a beautiful wife and lovely children, and he was contented. Contented because he was wealthy, and wealthy because he was contented. And one day there visited this Al Hafed an ancient priest, and that priest sat down before the fire and told him how diamonds were made, and said the old priest, 'If you had a diamond the size of your thumb you could purchase a dozen farms like this, and if you had a handful you could purchase the whole county.'
>
> "Al Hafed was at once a poor man; he had not lost anything, he was poor because he was discontented, and he was discontented because he thought he was poor."
>
> - Russell H. Conwell, 'Acres of Diamonds'[1]

1. The Nature of Preference Intensity

In the last chapter we canvassed two proposals for performing direct comparisons of individuals' utilities and found them wanting. Our discussion of these proposals did not conclusively demonstrate that direct interpersonal comparisons are impossible; despite the failure of the two accounts scrutinized above, the advocate of direct comparisons may yet cling to the hope that another account will be forthcoming, perhaps couched in more refined behavioural or physiological terms, which will succeed. The hope must be slim, however. The proposals which we

---

[1]Cleveland, n.p., 1905. Reprinted in Scott H. Partridge, Cases in Business and Society, 2nd ed. (Englewood Cliffs, N.J.: Prentice Hall Inc., 1989), pp. 33-40.

have so far examined are to my knowledge all and only those which have so far been advanced in any serious detail, and hence their failure provides some grounds for thinking that direct interpersonal comparisons are faulty in principle. The present section is devoted to confirming this suspicion.

Our goal in scrutinizing behavioural and physiological comparisons above was not merely to demonstrate that they fail. They certainly do fail, but more importantly they fail in an instructive manner. Recall that interval rankings provide us with complete information about the relative strengths of each individual's preferences. The proponent of direct interpersonal comparisons holds that there is something more to preferences than what is given by such rankings; that preferences have "absolute" magnitudes, not merely magnitudes in relation to other f an individual's preferences, and hence that additional information is required beyond that necessary for constructing personal interval scales in order to secure interpersonal comparability. Metaphysically speaking, the working assumption of direct interpersonal comparisons is that the intensity of a given preference is a monadic property o that preference, one which it possesses independently of the strengths of the other preferences with which it resides in someone's preference set. On this assumption, it makes perfectly good sense to look for ways of directly

comparing the intensity of one or more of an individual's preferences with one or more of someone else's, without taking any cognizance of their remaining preferences. Yet in examining proposals for performing direct comparisons, we arrived at the conclusion that a given bit of behaviour or physiology could be construed as evidence for the strength of a certain preference or preferences, only if we were prepared to make assumptions concerning the intensity of other preferences not originally under consideration.

These facts require an explanation. The one I shall defend is that direct interpersonal comparisons fail because preference intensities are not monadic properties at all, but are rather irreducibly relational properties of the preferences which possess them.

This is hardly a transparent claim, and making it out is going to take some effort. As a first step towards clarification, my main contention in what follows will be that preference intensities should be construed in the first instance as relations between an individual's preferences, relations whose sole purpose is to guide the apportionment of resources available to her for the satisfaction of those preferences. It is of course possible to view any relation from either "end", so to speak, as a property of one of its relata. E.g., our planet has the "property" of being about 500 light-seconds away from the Sun; equivalently, the Sun is about 500 light-seconds from

Earth. It is relatively uncontentious, I take it, that these so-called properties are relational ones; strictly speaking the distance between the Earth and the Sun is not a property of either, but a relation between the two. Similarly (no doubt this is a little more contentious), preference strengths are properly regarded as relations between a person's preferences. There is no real harm done in viewing them as properties of individual preferences, just as viewing the distance between the Earth and the Sun as a property of either is a harmless enough diversion. No harm, that is, so long as we are careful to remind ourselves that preference intensities are relational properties, and not monadic properties possessed by preferences independently of their relation to the other preferences of a particular individual.

I shall be attempting to elucidate and provide support for this relational view primarily by means of an analogy drawn between preference strengths and the priorities of tasks running on a multitasking computer. Before proceeding to that analogy, it may be helpful to distinguish the view from a couple of other claims with which it might be confused or on which it might be thought to depend.

We had occasion in the previous chapter to remark on certain respects in which value is a relational concept. For one thing, value is a relation between persons and objects, events, or states of affairs: there is no value

without valuers. We further noticed that when fully cashed out, value is not a simple dyadic relation holding between individuals and objects, events, or states, but rather a triadic relation: strictly speaking people do not value outcomes singly, they value them only in relation to other outcomes. A fortiori it makes no sense to ask how much an individual values outcomes taken singly. This is not what I have in mind in advancing the present relational view, however; the claim that preference intensities are irreducibly relational is intended to go well beyond the fact that value is properly construed as a three-part relation, and hence that it makes no sense to ask how much individuals value some outcomes except in relation to others. This should in any case be obvious enough once we shift from talk of individuals' values to talk of their preferences. It's meaningless to speak of the degree of A's preference for x and leave it at that, just because it's not possible for A to have a preference for x simpliciter. But granted that preferences are relational in this sense, my claim is the stronger one that preference intensities are relational. Not only is it incoherent to speak of the magnitude of A's preference for x, except in relation to something else y; ultimately it is also incoherent to speak of the magnitude of A's preference for x over y, except in relation to her other preferences, e.g. her preference for z over w.

Secondly, it should be emphasized that the thesis I am advancing is a metaphysical one, a thesis about the nature of preference intensity, and not (in the first instance) a thesis concerning the measurement of preference intensities. Measurement is in one sense inevitably relational: no magnitude of any sort can be specified except in relation to the magnitudes of other things. An object's length, e.g., must be specified in terms of the length of something else, whether that something else happens to be a conventional standard of measurement in the domain at issue (a standard meter, say) or a more arbitrarily selected reference point.[2] This is a fact about measurement, however, and not a fact about the nature of the attributes to be measured. Lengths are commonly supposed to be monadic properties, despite the fact that they can only be specified in relation to the lengths of other objects. (Some have held that the length of an object should be analyzed

---

[2]Measurement theory reflects the fact that all measurement is relational by explicitly treating all magnitudes as relations. In treating the measurement of length, e.g., the measurement theorist begins not by supposing that the objects in a set to be measured possess a certain property (i.e., 'length'), but rather that they are ordered by a certain relation: namely, the relation 'longer than'. If this relation is assumed to weakly order the objects in the set, then it can be shown that the "empirical structure" comprised of the objects together with the hypothesized relation can be consistently represented by an ordinal scale. Further conditions imposed on the structure suffice to guarantee constructability of a ratio scale, one which is unique up to the choice of a unit of measurement, which is just the result one would expect for measurement of lengths; for details see Krantz et. al., Foundations of Measurement, Ch. 1-3.

in terms of the distance between its endpoints; but this is beside the present point, which has to do with whether the length of an object is a feature of that object itself, perhaps one possessed in virtue of relations which its parts bear to one other, or whether it's something which the object possesses in virtue of relations to other wholly distinct objects.)[3]

---

[3]The exact content of the supposition that lengths are monadic properties is not easy to specify. Intuitively a monadic or "intrinsic" property is one which does not depend in any way on other things. Jaegwon Kim has attempted to capture the intuition by defining relational or extrinsic properties (he calls them "external" properties) as ones the possession of which entails the existence of at least one other wholly distinct object. ("Psychophysical Supervenience", Philosophical Studies 41 (1982), pp. 51-70.) David Lewis has noted a difficulty for the account, however. Define the complementary properties of 'accompaniment' and 'loneliness' as follows: something x is accompanied if and only if it coexists with at least one other wholly distinct object, and lonely if and only if it does not. Then Kim's idea is that extrinsic properties are ones which imply accompaniment, whereas intrinsic properties are compatible with loneliness. (One property implies another if and only if necessarily, anything which has the former has the latter.) But now consider the property of loneliness itself: it is just as extrinsic as accompaniment is, yet obviously it does not imply accompaniment, and it is clearly compatible with being lonely. ("Extrinsic Properties", Philosophical Studies 44 (1983), pp. 197-200.)

This particular difficulty can be met by modifying Kim's definition so that an extrinsic property is one such that either it or its negation implies accompaniment. (Note that in general the negation or complement of any property which implies accompaniment will itself be an extrinsic property which does not imply accompaniment, on pain of rendering loneliness logically impossible.) Other extrinsic properties are not so easily dealt with, however. E.g. 'being the fattest pig' is compatible with loneliness, but so is its negation (since something can fail to be the fattest pig in virtue of not being a pig at all). One possible solution may be to invoke a primitive notion of "natural" properties/relations, and maintain that Kim's original definition works as intended for this class. (Cf.

Like lengths, degrees of preference are specifiable only in terms of the strengths of other preferences. In contrast with lengths however, what I want to claim is that the intensity of a given preference is not an attribute possessed by that preference independently of relations it

---

Lewis, On the Plurality of Worlds, Oxford: Basil Blackwell Ltd., 1986, pp. 61ff.; unfortunately Lewis relies here on a primitive distinction between natural properties and natural relations, thus begging the question of how to distinguish intrinsic from extrinsic properties within the class of natural attributes. Following David Armstrong, we might characterize natural attributes as those referred to by a completed science; on reasons for thinking that negative and disjunctive properties will not be among them, see his Universals and Scientific Realism Vol. II, Cambridge University Press, 1978, Ch. 14.)

Apart from the issue of how monadic properties in general should be characterized, lengths and other properties which admit of degree present special difficulties of their own. Lengths are presumably natural enough; but could an object have just the length which it does if nothing else existed? It certainly seems to make sense to suppose that an object could be extended even though nothing else existed; but positivistically minded philosophers are inclined to answer the question in the negative, on grounds that operationally speaking the assignment of a determinate length to an object depends on the existence of other objects to which its length can be compared. We can avoid these more abstruse questions regarding the metaphysical status of lengths by noting that although it's true that an object's length can only be specified in terms of the lengths of other things, nevertheless something's having the length which it does does not depend on its having a certain length relative to any particular object. Consider any three wholly distinct objects a, b, and c: a could be longer than it is in relation to b; but a could also be longer in relation to c without being longer in relation to b (i.e., both a and b could be longer in relation to c). It is this possibility of variation independently of any particular object which for present purposes may be taken to mark lengths as monadic properties. Whether there is at base any real difference between something's increasing in length v. everything else's decreasing in length, whether it's possible for everything to have doubled in size overnight, etc., are questions we pass over.

bears to an individual's other preferences. In this respect preference intensities are rather like proportions. To attribute a proportional magnitude to something is not merely to describe it in terms of the features of something else; it is to commit oneself (logically) to the existence of another thing to which it stands in an inverse relationship. To say that a portion of a cake is 1/4 of the whole, e.g., is to say that it bears a certain relation to the size of the remaining cake; and moreover that an increase in its proportional size would entail a corresponding decrease in the proportional size of the remainder.

Preference intensities, I believe, operate in very much this fashion. They differ in one extremely important respect from most other proportional magnitudes, however. I have suggested not only that preference strengths are relational in character, but that they are "irreducibly" relational; and it can hardly be maintained that proportional magnitudes are generally thus. The parts of a cake (and just about anything else one cares to mention) bear the proportional relationships which they do to each other because each of them possesses a specific weight or volume. Since the weight or volume of each portion is a monadic property of that portion, not a relational one, at least proportional magnitudes of cake parts ought not to be construed as "irreducibly" relational.

Preferences are not cake parts, however, and whatever

the fate of the latter, I think that a good case can be made for saying that magnitudes of the former are not only relational, but irreducibly so. The key to understanding the nature of preference intensity, and ultimately to unlocking the problem of interpersonal comparisons, lies in arriving at a sound conception of the role that preference intensities may be understood to play in individuals' psychologies. Towards that end, consider the following analogy.

Imagine that we have before us a computer containing a single processing unit, and running under a "multitasking" operating system. Since the computer contains only a single processor, it can strictly speaking execute only one task at a time. The computer's operating system can create the illusion of two or more tasks being executed at once, however, through a process known as "time-slicing". Very roughly, what the operating system does is allocate a small "slice" of processor time (typically on the order of a few hundredths or thousandths of a second) to one of the tasks, then put that task on hold while it executes the next one, and so on. Though at any given instant the processor is executing only a single task, if the operating system rotates the various tasks quickly enough they appear to be running simultaneously.[4]

---

[4]I have chosen to focus on a single-processor computer simply because the concept of time-slicing is particularly easy to grasp with respect to such a machine. Nothing of

Suppose now that we set our computer to executing a simple program; a program, say, which will calculate pi to a few thousand significant digits. The program is intended to serve here as the analog of a preference or set of preferences; it is what guides the machine's overt behaviour, that behaviour visible to the user. Speaking somewhat fancifully, we may say that the program governs what "choices" the computer will make from among available alternatives (rather simple choices, to be sure, given that no matter what situation we put it in, the computer will go on calculating pi if that alternative is available, and otherwise do nothing of interest).[5]

At this stage there would appear to be little grounds for attributing anything like the analog of a preference intensity to the task which the computer is executing. It might be supposed that the strength of our preference-analog is to be discovered in the rate at which the computer calculates pi or some such. But this suggestion

---

importance depends on the choice; we could as well have made do, at the cost of irrelevant complication, with a multiple-processor machine.

[5]Perhaps it would be more accurate to say that the program serves as the analog of a desire, or preference for some state of affairs over the status quo (see note 8 of the previous chapter). It seems doubtful that computers possess anything analogous to full-blown preferences, including differential "attitudes" towards pairs of outcomes neither of which are or will be actual.

Henceforth I shall drop the scare quotes in speaking of the computer's "choices", "attitudes" and the like. Those "worried" about the dangers inherent in anthropomorphizing a mere machine may sprinkle in imaginary ones in liberal quantity.

seems to confuse the intensity of the computer's preferences with the resources it has available for satisfying them. If we were to decrease the processor's clock speed (the speed at which it actually manipulates data), the calculation of pi would be correspondingly slower. But the program itself would not have changed in any way; the computer's resources for executing the task would merely have decreased. Similarly, I think we want to say, it is in principle possible for two people to have exactly similar preferences and preference strengths, yet differ in the resources they have at their disposal for satisfying them. If someone is a more efficient satisfier of preferences in virtue of possessing superior mental or physical abilities, e.g., this shouldn't automatically lead us to conclude that she possesses stronger preferences. Likewise we need to distinguish between the intensity of the computer's tasks and the rate of their execution.

Now suppose that we run another program on the computer, so that it is sharing its processing time equally between two tasks. In this case it may make sense to ascribe intensities to the tasks being executed, at least in relation to one another. Since the computer is sharing its time equally between the two tasks, and will do so in any situation in which it has the opportunity to execute both of them, it seems plausible to suggest that whatever the intensity of the first, the second has exactly the same

intensity. Note that assignment of equal intensities to the two tasks is consistent with the proviso that we distinguish strengths of preference from the resources available for satisfying them. Were we to increase or decrease the speed of the processor, the rate of execution of the tasks would increase or decrease by the same amount, so that the assignment of equal intensities to them holds up.

Let's introduce one more complication. Most multi-tasking operating systems provide a means of specifying a priority for each task being executed, which they use to gauge the number of slices of processing time that a given task will receive before it is put on hold and the processor is given over to the next task in the rotation. Suppose that our operating system provides for 10 levels of priority, linearly ordered, so that a task which is assigned a priority of 10 receives ten times as much processing time as a task with priority 1. Suppose further that we assign a priority of 4 to the first program which we ran, and 2 to the second. Under these circumstances, the computer will devote roughly 2/3 of every second of its processing time to the first task, and 1/3 to the second task.

We now have very clear grounds, I think, for assigning analogs of preference intensities to the two tasks. We can judge quite safely that the intensity of the first task is

twice that of the second, since in any given situation the computer will devote twice as much time to executing the former as the latter. This holds for variations in processor speed, of course: vary the clock speed how we will, the computer will still devote twice as much time to the first task as to the second. It also holds in situations where we set the computer to executing further tasks. Were we to start up another process with a priority of 2, e.g., then 1/2 of the available processing time would be devoted to the first task, and 1/4 of the time to each of the subsequent two. Realistically the management of each concurrent task involves a certain amount of system overhead, and hence the total amount of processing time available for executing tasks collectively will decrease as additional tasks are run. Nevertheless, no matter how many processes we run (within the limits imposed by the operating system) and no matter what their priority levels, the amount of processing time devoted to the first task will always be roughly 2:1 in relation to the second.

Having developed the analogy to the point where fairly clear sense can be made of attributing intensities to the tasks which the computer is executing, a couple of points need to be emphasized. The first is that the actual numbers used to assign priorities to the tasks on our hypothetical computer have precisely no significance in themselves. We might just as easily have designed an

operating system (exactly the same operating system, in all relevant respects) which interpreted priority levels of '2' and '4' to mean that any task with the former priority was to receive twice as much processing time as one with the latter, instead of vice versa. Indeed, we might just as easily have chosen random letters from the Hebrew alphabet to represent priority levels, were it not for the inconvenience of trying to remember which letter stood for which priority. From the computer's point of view, whatever representations we have settled on are merely handy labels to which it attaches some predetermined significance in figuring out how to allocate its limited processor time. Thus what justifies the assertion that the first task we ran has twice the intensity of the second is not that the numeral '4' on its usual interpretation refers to a number which is twice '2' on its; but rather the fact that the operating system interprets the labels '4' and '2' to mean that any task bearing the former should receive twice as much processing time as one bearing the latter.

The second point is related, but runs deeper: Given that it is the way in which the operating system translates the symbols which we have chosen to represent priority levels into behaviour with respect to its various tasks which justifies claims regarding the intensities of those tasks, it is essential to realize that the system interprets priority levels in an <u>exclusively</u> relational fashion.

By itself a priority of 4 or anything else assigned to a task has no meaning at all, no implications of any sort for the machine's behaviour; it only acquires significance in relation to the priorities of other tasks (which themselves have only relational significance). If the machine is currently executing only one task, then that task will receive all of the available processor time, regardless of what priority it has been assigned. If additional tasks are executed, then the amount of processing time devoted to the original will be determined by the significance which the operating system attaches to its priority in relation to the priorities of all the other tasks (again, regardless of what priority the original task was assigned). An upshot of this is that apart from differences in the (arbitrary) internal representation of task priorities, there is no difference between the machine executing a pair of tasks with priorities of 4 and 2 and executing the same pair of tasks with priorities of 2 and 1, e.g.; in respect of all relevant behaviour, internal as well as external, the situations are functionally indistinguishable.[6]

---

[6]There is of course the following nominal difference between the two situations: If we were to start up another process with priority 6, then in the first situation the original two processes would collectively consume one half of the available processor time; whereas in the second case the original two processes would consume one third of the total time. But to view this as a relevant difference would be to mistakenly suppose that a priority of 6 assigned to the third task has some special significance in abstraction from the priorities of the tasks already being executed. The appropriate question to ask is rather

The implications of interpreting priority levels in this relational fashion may be illuminated by contrasting the system so far described with one that behaves rather differently. Suppose that rather than allocating processing time as described above, our operating system interpreted a priority level of 10 to mean that any task with that priority was to receive exactly 10 out of each 100 slices of processor time. It would not be at all difficult to design and implement such an system, and for all I know someone may already have done so. Note however that a system of this sort would be considerably less flexible in operation than the one originally described. For one thing, if only a single task was being executed, the machine would be sitting idle at least 90% of the time. This problem could be circumvented by adding higher priority levels, but other inflexibilities in the design could

---

whether there would be any significant difference between the machine executing tasks with priorities 4-2-6 and executing the same tasks with priorities 2-1-3; and the answer is none at all.

Note that while there are certain combinations of tasks and priorities whose functional upshot can only be realized in one way (the only way for one task to receive ten times as much processing time as another one, e.g., is to assign them priorities of 10 and 1 respectively), this is an artifact of our simple analogy and cannot be generalized. There is no reason in principle why a multitasking operating system might not provide for an infinite number of priority levels (indeed, an infinite number of priority levels might conceivably be easier to implement than a finite number on an analog machine). In this case, given any specification of tasks and priorities, it would always be possible to describe a functionally equivalent situation in which the machine is executing the same tasks with different priorities.

not be so easily worked around. In general, the efficient execution of any set of tasks would involve the operator having to carefully match the priorities of tasks to each other in order to ensure that the processor was not sitting idle some percentage of the time. But then any addition or subtraction of tasks would involve performing the match all over again, since it would be impossible to add tasks without first decreasing the priorities of some or all of the tasks currently being executed, while deleting tasks would result in squandered processor time unless the priorities of some or all of the remaining ones were increased. The beauty of interpreting priority levels in a purely relational fashion is that none of this is necessary; whatever tasks are being executed and whatever their priorities, the system will automatically take full advantage of the available processing time, and will continue to do so whether or not processes are added or deleted.

Hence the attachment of an exclusively relational significance to priority levels by our original operating system is of direct utility in maintaining flexibility as regards the number and priority of tasks being executed, while at the same time guaranteeing the efficient use of available resources. The price of this guarantee (really it is no price at all) is that the intensities of the computer's tasks can only be described as being "irreducibly relational". It is not possible to identify the

intensity of a given task with some fixed amount of processing time, since in contrast with the operating system described in the previous paragraph, there is no such thing as the amount of processing time that a task will receive in abstraction from other tasks. It is of course possible to state the amount of processing time which any pair of tasks will receive in relative terms (a task with priority 4 will always receive twice as much time as one with priority 2, e.g.); but this is compatible with the pair of tasks receiving all of the available processing time, or virtually none of it. If we are interested in a more concrete specification of task intensities, the best that we can possibly do is to state what proportion of processor time each task will receive within a fully specified system of tasks and priorities; for it is only within the context of such a fully specified system that any task will receive a determinate amount of processing time. What makes task intensities on our multitasking computer irreducibly relational is precisely the fact that priority levels do not have any specific implications for the machine's behaviour apart from such a fully specified system.

Persons are not computers (at least not simple-minded ones of the sort we have been contemplating), and it would be rash to draw any substantive conclusions directly from this simple analogy. One of the more obvious shortcomings of the analogy is that multitasking computers typically

allocate processor time without regard for the amount of time that will actually be required to complete a given task. A more sophisticated system might be expected to provide a facility for differentiating tasks not only with respect to the importance of their completion, but with respect to the importance of their completion within some specified interval; such a system would serve as a better analogy in the present context, since complicated organisms like ourselves may obviously attach some importance to achieving ends at a certain time or within a certain period of time, and not merely at some time in the indefinite future. Another deficiency of the analogy is that whereas computers possess only a single resource which they can differentially allocate towards completing their assigned tasks,[7] we have at our disposal a considerably wider range of resources for pursuing our ends. These may be taken to include not only "internal" resources like mental and physical capacities, but also a tremendous variety of external resources; anything which may be of service in satisfying preferences is a candidate here, including so-called natural resources, tools or other artifacts which can be used in conjunction with our native capacities, money or more substantial goods with which to barter, etc.

---

[7]I am ignoring here other system resources, such as memory, disk and tape storage, and other input/output devices, which are typically allocated to processes on a first-come, first-serve basis.

It is not difficult to think of other respects in which the analogy is not as tight as it could be. We might endeavor to complicate the story in various ways in an effort to make it more realistic, but I shall not attempt to do so. I do not in any case wish to pin a great deal of weight on the analogy; it is intended merely to be suggestive of the general manner in which preference intensities may be understood to function in our psychologies. In exploring proposals for directly comparing the strengths of different individuals' preferences, we noticed that behavioural or physiological data could be counted as evidence for the strength of someone's preferences only if we were prepared to make assumptions concerning the strengths of her other preferences. The most straightforward explanation for this fact, though one which is not in itself very illuminating, is that an individual's preferences simply do not have intensities on their own, but only in relation to one another. Despite its obvious shortcomings, the analogy sketched above can provide some badly needed content for the explanation, by helping us to see how it could come to pass that preferences don't have intensities in abstraction from the other preferences with which they reside in an individual's preference set. Preference intensities may be essentially or irreducibly relational because their sole function is to guide the apportionment of whatever resources are available to a person for pursuing her ends, just

as the sole function of priority levels on a multitasking computer is to guide the allocation of processing time to different tasks.

"The motive in one mind," Jevons wrote, "is weighed only against other motives in the same mind, never against the motives in other minds." Just so. Jevons did not go on to address what in retrospect appears to be the obvious follow-up question, however: What possible use would creatures such as ourselves have for preference intensities which somehow outstripped a determination of the strengths of each of our motives in relation to all of the others? One of the virtues of the computer analogy is that it invites us to take up the "design stance" towards ourselves and confront this issue head-on. I do not think that it is lack of imagination on my part which suggests that preference intensities which possessed some significance beyond their significance in relation to each other would have no part to play in the life of the organism. So far as I can see, the only plausible role for preference strengths to play is in determining how diverse and variable amounts of resources will be directed towards satisfying various preferences; and for this role, preference strengths which had more than a relational significance would be of no use at all.

Notice that there is no reason in principle why we couldn't be built in such a way that the mechanisms which

serve to direct our internal and external behaviour had some sort of non-relative magnitude associated with them which determined that fixed amounts of certain resources would always be directed towards certain ends. Indeed, to some extent we are built like that. The mechanisms which control our heartbeat and body temperature, e.g., seem to work in roughly this fashion. These mechanisms are not preferences, however. I do of course prefer that my heart go on beating to its stopping any time soon. But it is not this preference which causes my heart to continue beating; were I to cease to have the preference my heart would continue to beat nonetheless, unless I undertook drastic measures to interrupt the causal processes involved.

To have non-relational magnitudes associated with our motives proper -- i.e., with the determinants of voluntary behaviour, as opposed to the involuntary mechanisms which maintain critical aspects of our systems as a precondition for engaging in voluntary behaviour -- would be as unduly limiting as designing a multitasking operating system which devoted fixed amounts of processing time to specific tasks, regardless of what other tasks were being executed and what their priorities were. The ends towards which our behaviour is directed are highly variable, not to mention the importance which we attach to those ends in relation to one another; and the resources available for pursuing our ends also vary widely both in degree and kind. If it were

genetically determined that specific amounts of certain resources would always be directed towards certain behaviours, this just wouldn't allow for the kind of flexibility in goals and resource use which we may presume has been highly instrumental in allowing our species to flourish in a multitude of environments. "Much better to design a system," we can imagine Mother Nature saying to herself, "that can direct its behaviour towards any number of unspecified ends, and will automatically make full use of whatever resources are handy for pursuing those ends."

My suggestion then, is that the sole function of preference intensities in our psychologies is to determine how available resources will be directed towards satisfying the preferences to which they attach, in the context of variable other preferences with variable strengths of their own. To know something of an individual's preferences is to know what ends her behaviour will be directed towards in various circumstances. To know what the intensities of her preferences are is to know something about how she will allocate the limited resources at her disposal in pursuing those ends.[8] Preference intensities, in other words, serve to regulate behaviour within an integrated system of

---

[8]Including how she will choose among risky alternatives. As the discussion in Section 2 of the preceding chapter should have made clear, risk-based measurement of individual utilities essentially involves giving an individual some outcome for certain, and then seeing how she will make use of this resource in pursuing other ends.

preferences. In abstraction from such a system they don't do anything at all, just as the priority levels of tasks on a multitasking computer have no specific implications for behaviour outside of a fully specified system of tasks and priorities.

The case for this relational view of preference intensity is not yet complete. I have offered it primarily as a means of explaining the observed failure of behavioural and physiological utility comparisons. As with all abductive inference or "inference to the best explanation", however, the full worth of this particular explanation can only be judged on the basis of how well it coheres with our judgments within a wider explanatory framework. The invitation to take up the design stance above and consider what use we could have for preference intensities which had more than a relational significance served to locate the relational view within a somewhat wider framework -- namely, the framework of evolutionary biology -- and within this context the explanation holds up pretty well. But there are other considerations of immediate and obvious relevance here. In particular, we have yet to consider how well the relational view meshes with judgments regarding individual welfare; with our untutored views on interpersonal utility comparisons; and ultimately, with our "considered moral judgments", especially as regards the distribution of goods.

I think we shall find that the relational view fits very happily within this wider framework, but that must remain a promissory note for a while yet. In the meantime, I hope that enough has been said to cast very serious doubt on the possibility of performing direct interpersonal comparisons of any sort. The problem with direct utility comparisons is that by definition they try to fix the strengths of some of an individual's preferences in the absence of information about her remaining preferences, information which is critical if the hypothesis I have been defending is substantially correct. If preference strengths really are irreducibly relational, then attempting to directly compare utilities across individuals is a bit like trying to figure out which of two spatial locations is farther away; but not farther from one's present location, or indeed from any location at all.

2. The Nature of Welfare

The problem of interpersonal comparisons is commonly posed as the question of how to define an objective "unit of comparison" by means of which to commensurate the interval utility scales of different individuals.[9] In the

---

[9]Thus Harsanyi, remarking on his behavioural comparisons, writes: "Of course, when we first define a von Neumann-Morgenstern utility function for each individual in the usual way, we shall normally choose an independent utility unit for each individual. But, then, we must engage in interpersonal utility comparisons in order to estimate conversion ratios between the different indiv-

last section, however, we discovered reasons for thinking that an individual's preferences do not have strengths except in relation to her other preferences; and interval scales of utility convey full information about the relative strengths of an individual's preferences. Have we not therefore discovered grounds for thinking that there is no such thing as an objective unit of comparison, and hence for dismissing the possibility of objective interpersonal comparisons once and for all?

The conclusion would be hasty. There is far too much at stake to accept it straight away, without fully exploring the alternatives. The stakes are obviously high for the utilitarian, whose doctrine is largely contentless in the absence of some form of interpersonal comparisons. But it is not only the utilitarian who has an interest in the current proceedings. No halfway sensible moral theory can completely avoid questions of aggregate welfare in making policy recommendations. Rawls' theory of justice attempts to side-step problems regarding the measurement and compar-

---

iduals' utility units." ("Nonlinear Social Welfare Functions: Do Welfare Economists Have a Special Exemption from Bayesian Rationality?", Theory and Decision 6 (1975), pp. 311-32; reprinted in Essays on Ethics, Social Behaviour, and Scientific Explanation, pp. 64-85.) In a similar vein Nozick writes: "The problem of interpersonal comparisons is to specify, for two persons' utility functions which are given, a common unit and a common zero point, to calibrate their utility functions so that the unit on each represents the same difference of utility for each, and the zero point on each represents the same degree of wanting." ("Interpersonal Utility Theory", Social Choice and Welfare 2 (1985), pp. 161-79.)

ison of welfare by appealing to an index of "primary goods", goods which everyone is presumed to want whatever their plan of life. But there is no plausible way of constructing the index short of considering to what extent different people value different primary goods; the doctrine of primary goods does not so much side-step problems of intercomparability as paper over them.[10] Even Nozick seems prepared to admit that it might sometimes be permissible to violate individual rights in order to avoid a great catastrophe.[11] In view of the importance of the matter for ethical theorizing in general, it behooves us to take a closer look at the problem of interpersonal comparisons in light of our recent results.

If the question is what more must be added to personal interval rankings in order to secure an objective unit of comparison, then the answer of the previous section is that there cannot be anything more; that given the relational nature of preference intensity, once we have in hand

---

[10]Cf. Arrow, "Some Ordinalist-Utilitarian Notes on Rawls's Theory of Justice".

[11]Anarchy, State and Utopia, p. 30n. Great catastrophes presumably depend for their greatness on summing the not-so-great (from an impersonal point of view) catastrophes of numerous individuals. Nozick actually refers to the avoidance of "catastrophic moral horror", so it may be that he has in mind only that individual rights might be violated in order to avoid very large-scale violations of rights, on the order of the atrocities committed by Hitler or Stalin. I leave it to the reader to judge whether a theory which didn't permit violations of rights in order to prevent other sorts of calamities (large-scale natural catastrophes, say) could be labelled "halfway sensible".

exhaustive interval scales of utility we have all of the information about individuals' preferences and preference strengths which it is possible to have. Thus if objective interpersonal comparisons are possible at all, they must be possible on the basis of personal interval rankings. But interval rankings merely provide us with information about the relative strengths of each individuals' preferences. How then is any objective comparison of preference strengths across individuals possible?

The answer is that exhaustive interval rankings do not merely provide us with information about the strengths of an individual's preferences in relation to her other preferences. Or rather, while such rankings do simply provide information about the relative strengths of preferences, the fact that they are exhaustive allows us to say considerably more than what is expressed in statements of the form "A's preference for x over y is so many times as strong as her preference for z over w." The analogy of the previous section may prove instructive here. Task intensities on our multitasking computer were irreducibly relational, we noted, in virtue of the fact that the operating system attaches a purely relational significance to priority levels. But it did not follow that we were limited to expressing task intensities in purely relative terms: "This task has so many times the intensity of that one." Such relative specifications of task intensities are

indeterminate to the extent that one task's having a certain intensity relative to another one is compatible with the pair of them receiving all of the available processing time, or virtually none of it. We noted that a more informative characterization of task intensities was possible, however, in the context of a fully specified system of tasks and priorities; for within such a context we can state exactly what proportion of the available processing time will be devoted to each task.

Similarly, it should be possible on the basis of complete interval rankings to identify how much weight a given preference carries within an individual's system of preferences as a whole. Here we face a small difficulty, however; namely, how to characterize an individual's system of preferences as a whole. What is to count as a "complete" or "exhaustive" interval ranking? The problem is that preferences are individuated on the basis of outcomes, and we have so far been following received practice in saying very little about the nature of these outcomes. Outcomes must be mutually exclusive in order to be fit objects of preference at all; but beyond this, virtually anything seems to go.[12] Sometimes outcomes are understood

---

[12]The mutual exclusivity requirement is often met implicitly rather than explicitly in ordinary speech. If it makes sense to say that I prefer having an apple and a banana to having an apple, e.g., this is only because it is understood that the latter outcome includes my not having the banana. (Actually, considerably more is implied than my simply not having the banana. Presumably I don't prefer

simply to be objects (or more generally, congeries of objects -- "bundles of goods"), particularly within micro-economic contexts. But just as often they are conceived of as possible states of affairs of greater or lesser extent; or, when temporal aspects are relevant, as datable events.

Given the bewildering variety of items over which a person may be said to have preferences, it makes little sense to suppose that we could list off a person's preferences, say "these are all there are", and proceed to calculate the proportionate weight of each one on the list. More to the point, a person's unrestricted preferences overlap in ways which make them unsuitable for present purposes. What we are looking for is a means of characterizing someone's overall system of values so that we can determine what weight a given preference carries within this system. Suppose that I prefer being a philosopher to being a lawyer. If this particular preference is to be granted a place within my overall system of values, then there are a host of other preferences which should not be counted along with it. For example, if I prefer being a philosopher and playing darts in my spare time to being a

---

having an apple and a banana to having an apple and not a banana but a million dollars instead, and the original statement of my preference wouldn't be taken to imply that I did. What seems to be implicit in the original statement, in addition to my not having the banana under the second outcome, is a kind of ceteris paribus condition: I prefer the apple and the banana to the apple alone, amounts of money in my possession as well as everything else held constant.)

lawyer and playing darts in my spare time, this preference shouldn't be granted a place in my system of values alongside the previous one; that would be a form of double-counting which resulted in a misleading picture of what I value.

Thus it won't do to include any old preference which a person might be said to have within her overall system of preferences; we need to be somewhat more precise in our characterization of outcomes than we have so far been. We can solve the problem of double-counting by limiting our attention to preferences over a set of outcomes each of which is mutually exclusive. (Being a philosopher and being a dart-playing philosopher are not mutually exclusive, e.g., so at least one of these two outcomes will not be in the set.) Not just any old set of mutually exclusive outcomes will do, however. We want a means of characterizing individuals' overall systems of value, and hence we must insure that the set of outcomes is rich enough so that anything which a person might care about will somehow figure in it.

This dual requirement of insuring that outcomes are mutually exclusive and at the same time rich enough to encompass everything that individuals care about can be met by taking outcomes to be entire possible worlds, and that is what we shall proceed to do. In doing so we are no doubt casting our net somewhat too wide. The structures in

our brains which subserve preference and choice presumably operate with respect to something less than maximal states of affairs. It is quite plausible to suppose that I can represent to myself being a philosopher and being a lawyer, e.g., and think the former preferable to the latter; but it is not in the least plausible to suppose that I can represent to myself entire possible worlds in any but the sketchiest of detail. ("Imagine a world pretty much like this one except that....") In taking outcomes to be possible worlds, however, we are not supposing that they figure in any direct way in individual decision-making. Rather, individuals' preferences regarding more limited states or aspects of the world can be understood as determining an ordering over possible worlds. People who prefer being a philosopher to being a lawyer, e.g., prefer worlds in which they practice philosophy to worlds in which they practice law.[13] Those who in addition prefer playing darts to playing checkers in their spare time prefer worlds in which they are dart-playing philosophers to worlds in which

---

[13] Every world in which they practice philosophy to any world in which they practice law? Presumably not; it would be a dedicated philosopher indeed who preferred philosophy to law at any cost. Rather, they prefer worlds in which they practice philosophy to worlds in which they practice law other things equal (cf. the previous note). There are important issues here having to do with what Peter Schotch has referred to as the "type-raising problem" -- i.e., the problem of raising the type of a preference relation, from a relation between possible worlds to a relation between propositions or sets of possible worlds -- which I shall for the most part do my best to avoid.

they are checker-playing philosophers, and worlds in which they are dart-playing lawyers to those in which they are checker-playing lawyers.[14]

In this fashion, we may suppose, individuals' preferences regarding various aspects of the world collectively serve to weakly order sets of possible worlds. The advantage of taking the objects of preference to be possible worlds themselves is that we do not risk leaving anything out of the picture so far as individuals' values are concerned, as we would if we attempted to construe outcomes more narrowly. If we have let in a bit too much, that simply means that possible worlds differ from each other in many inconsequential ways, ways which do not figure in anyone's system of values. Hence there will be large equivalence classes of worlds between which most individuals are indifferent (worlds which differ only in the exact number of stars in the Andromeda galaxy might belong to such equivalence classes, e.g., on the assumption that nobody much cares exactly how many there are); but this is a harmless enough result.

Limiting our attention to preferences over possible worlds will not by itself provide a means of characterizing

---

[14]What about their preferences regarding checker-playing philosophy worlds and dart-playing lawyer worlds? The ranking here plausibly depends on the relative strengths of their preferences for philosophy over law and darts over checkers: those who prefer philosophy to law more than darts to checkers should prefer checker-playing philosophy worlds to dart-playing lawyer ones.

individuals' overall systems of value. We also need to suppose that each individual possesses a most and a least preferred world (presumably, equivalence classes of most and least preferred worlds); or at any rate, that individuals' preferences tail off asymptotically at both ends if they do not actually arrive at most and least preferred worlds. This appears to be a reasonable enough assumption. I can think of lots of respects in which the world could be a better place from my personal point of view. But if I successively imagine various changes occurring for the better, at some point my catalogue of possible improvements begins to run out. Perhaps it's true that no matter how good things get, they could always be a bit better; but not much, I'm inclined to think, after a certain point. Similarly, if I imagine a series of incremental changes for the worse, at some point I begin to suspect that the world couldn't deteriorate much further; things could be different, all right, but they couldn't be much worse. Note that the supposition that individuals ideally possess maximally satisfying and maximally frustrating worlds does not imply that anyone's preferences could ever be fully satisfied. Marx attributed the movement of human history to the principle that as wants are satisfied, new ones are created.[15] If this principle is accurate then presumably a

---

[15] "The German Ideology" in Robert C. Tucker, ed., The Marx-Engels Reader, 2nd. edition (New York: W.W. Norton & Company, 1978), p. 156.

person could never be without unsatisfied preferences, since as her current preferences became progressively more satisfied she would proceed to acquire new ones. But this does not prevent us from roughly identifying maximally satisfying and maximally frustrating worlds as determined by her current preferences.

We now have sufficient resources to be able to characterize the weight which particular preferences carry within an individual's overall system of preferences. Let U be an interval utility function defined over possible worlds x,y,... for an individual A; and let m and l be a pair of most and least preferred worlds in the domain of U. The interval between m and l may be taken to represent A's total possible welfare; it is the difference between her being as well off and as poorly off as she could possibly be, according to her subjective assessment of all the factors which collectively determine her ranking of worlds. The interval between m and l has no significance in itself, of course, since its size depends on our having arbitrarily fixed the values of U for some pair of outcomes. It can however serve as the basis for calculating the weight of particular preferences within A's system of preferences: the weight of a preference may be identified with the proportion it represents of her total possible welfare. I.e., formally, the weight of a preference for x over y within A's system of values is given by:

(1) $$\frac{U(x) - U(y)}{U(m) - U(l)}.$$

Notice that this ratio is invariant with respect to positive linear transformations of U. Interval scales are so-called precisely because ratios of intervals do not vary with admissible transformations of the scale, and the formula above refers to just such a ratio of intervals.

Construing the strengths of an individual's preferences in terms of their proportionate contribution to her total possible welfare is, I believe, the best that we can possibly do by way of a concrete specification of preference intensities. And given that preference strengths so construed are invariant with respect to linear transformations, they are at least formally suitable for comparison across individuals. I do not wish to claim merely that this is the only determinate content that the notion of preference strengths can have given the relational nature of preference intensity, however. I think that this is how we do ordinarily conceive of welfare, both our own and others', though perhaps in a somewhat confused fashion. Most theorists have assumed, implicitly if not explicitly, that our everyday judgments of interpersonal welfare rest on comparing the absolute strengths of different individuals' preferences. Uncritical acceptance of this assumption is in my view more than anything else responsible for generating and sustaining the problem of interpersonal comparisons. If the assumption is correct, then obviously

preference strengths construed in terms of the ratio representation above cannot serve as the intuitive basis for our everyday judgments. The crucial question is whether the assumption is justified.

The easiest way to see that our ordinary judgments do not in fact depend on discerning the absolute strengths of preferences is to examine the notion of levels of welfare in the light of the proportionate interpretation of preference strengths sketched above. Strictly speaking utilitarianism does not require any comparison of welfare levels across individuals. In ranking policies, the utilitarian needs to know only the strengths of individuals' preferences regarding the outcomes of those policies. For a pair of outcomes x and y and individuals A and B, e.g., all the utilitarian needs to know is whether A prefers x to y more than or less than B prefers y to x; this is sufficient to determine which alternative will better promote aggregate welfare, regardless of what levels of welfare A and B would actually enjoy under the two outcomes. Nonetheless, we often do make judgments about levels of welfare -- about how well off someone is in the circumstances -- and certain features of those judgments weigh heavily in favor of construing the strengths of an individual's preferences in terms of the proportionate weight they carry within her system of values as a whole.

The level of welfare which an individual will enjoy at

an outcome is typically identified with the strength of her satisfied preferences at that outcome. This is how behavioural proposals for performing direct utility comparisons are supposed to work, e.g.: an individual's expressive behaviour is taken to indicate her overall degree of satisfaction with her situation, and then the strength of her satisfied preferences is supposedly inferred from her level of satisfaction. If preference intensities are irreducibly relational in nature, however, then welfare levels cannot be equated with the absolute intensities of individuals' satisfied preferences; they must rather be analyzed in terms of the proportionate intensities of satisfied preferences. Consider once again A's interval ranking of possible worlds, from her best world m to her worst one l. For a given outcome x, the interval between x and l may be taken to represent the strength of A's satisfied preferences at x. This interval has no significance on its own, any more than the interval between m and l does. The two intervals do however have significance in relation to each other: the ratio of the former to the latter states the extent of A's satisfied preferences at x in relation to her preferences in total. In other words, on a proportionate interpretation of preference intensities, A's level of welfare at an outcome x may be represented by:

(2) $$\frac{U(x) - U(l)}{U(m) - U(l)}.$$

Less tersely, A's level of welfare is fixed by the weight which her satisfied preferences carry within her overall system of value, by their proportionate contribution to her total possible welfare. Once again, welfare levels so construed are invariant with respect to linear transformations, and so are at least formally suitable for comparison across individuals.

It has been suggested to me that this is a wildly counterintuitive analysis of what it means for an individual to enjoy a certain level of welfare. My own intuitions suggest otherwise, though perhaps they are not to be trusted at this stage. I suspect, however, that my theory-laden intuitions are not really so radically at odds with the common sense view of the matter. A critical feature of the analysis just presented is that it makes the level of welfare which an individual will enjoy in given circumstances depend as much on the strength of her unsatisfied preferences as on the strength of her satisfied ones. Increasing the weight of A's unsatisfied preferences (in relation to her satisfied ones), e.g., would have the effect of increasing the denominator of the ratio above (or equivalently, decreasing the numerator), and hence of decreasing her level of utility. And this seems to be a pervasive feature of our common sense understanding of individual welfare. It was precisely the intuitive plausibility of the idea that levels of welfare depend as much

on unsatisfied preferences as on satisfied ones that we relied on in criticizing behavioural proposals for performing direct utility comparisons in the previous chapter. Behavioural proposals attempt to infer the strength of individuals' satisfied preferences from the level of satisfaction or dissatisfaction which they express with their situation. The problem with such proposals, we noted, is that there appears to be no way of choosing between the hypothesis that an individual who expresses dissatisfaction with her situation is miserable in virtue of the low strength of her satisfied preferences, or in virtue of the high strength of her unsatisfied ones. On the present analysis, there is no way of choosing between the two hypotheses simply because there is nothing to choose between them; given the relational nature of preference intensity, they express precisely the same underlying facts.

The short story about Al Hafed which prefaces the current chapter also plausibly depends on construing utility levels in terms of the proportionate contribution of individuals' satisfied preferences to their total possible welfares. After the visit from the priest Al Hafed is less well off than he was. But he has not lost anything. Nor, presumably, have the preferences on which his high level of utility prior to the priest's visit depended altered in any way with respect to each other.

Rather, he has acquired new, unsatisfied preferences; his satisfied preferences thereby account for somewhat less of the total than they previously did, and his level of welfare is therefore diminished.

I venture, then, that the proportionate interpretation of preference intensities and utility levels offered above meshes quite nicely with our intuitive understanding of individual welfare. When we judge that all things considered an individual is well off, our judgment has to do precisely with whether her own deepest concerns are met, and not at all with whether there is some absolute significance attaching to those concerns beyond the importance which she attaches to them in relation to her other concerns. Appealing to shared intuitions regarding welfare or anything else is a tricky business, however. While a common core of understanding is necessary in order for communication to be possible at all, there is always the risk that one's own intuitions have been shaped in ways that will lead them to diverge from those of other people in important underlying respects. Readers who remain unconvinced may perhaps try out the following thought experiment as a means of determining how far their intuitions depart from mine. Suppose, contrary to my hypothesis, that preference intensities do have some significance beyond their significance in relation to the other preferences with which they reside in individuals' preference

sets. Now ask yourself whether you would think the world a better place if the absolute strengths of your currently satisfied preferences increased substantially, on the assumption that the importance which you assign to things which are not already in your possession also increased by a similar magnitude. I.e., ask yourself whether it would matter to you if in Jevons' terms the strengths of all of your motives, unsatisfied as well as satisfied, increased a thousand-fold in every direction. I do not think it would matter to me.

There is one more crucial bit of evidence to be offered in support of the view I am defending: The proportionate interpretation of welfare guarantees that marginal utility will eventually diminish for all persons, with respect to all goods. It is important to note that there is nothing in the interpretation which insists that a specific increment of some good must be valued less than the previous increment; it is perfectly consistent with a proportionate construal of preference strengths that I might attach a constant or increasing subjective importance to increments of goods over some range. In terms of our possible worlds representation of preferences, movement from one world to another one in which I possess an extra increment of some good might account for 1% of my total possible welfare, movement to another world in which I possess an additional increment might account for 2% of my

possible welfare, and so on for a considerable ways. This is as it should be; as has often been noted, it is reasonable to suppose that people sometimes value the next increment of a good more highly than the last (e.g., in situations where a fixed amount of the good in question is required in order to secure something which a person wants badly, and the additional increments are sufficient to get over the threshold). If an increment of a certain good accounts for 1% of my possible welfare, however, then obviously subsequent increments cannot also count for 1% or more indefinitely; at some point my scale of value begins to run out, and additional increments must be prized less.[16]

Clear recognition of the principle of diminishing marginal utility was instrumental in the development of market theory by Jevons and others in the late nineteenth and early twentieth centuries. The principle has usually been viewed with something less than perfect equanimity, however. Although it played a pivotal role in deriving market equilibria, there did not appear to be any way of explaining it in terms of more basic facts, and theorists

---

[16]Note that the proportionate construal of welfare does not strictly speaking entail eventually diminishing marginal utility for all goods. It rather entails diminishing marginal utility for any good which an individual values to any appreciable extent. It remains theoretically possible for someone to value each of an infinite number of increments of some good as much as she values the first; but only if she attaches an infinitesimal value to each increment.

as diverse in time and preoccupation as Edgeworth, Pareto, and R.M. Hare[17] have tended to view the principle for lack of a better characterization as a wholly contingent affair; a kind of happy accident which for no apparently good reason applies to people and goods generally and makes economics possible.

One reason why economists in particular have not seen their way clear to explaining diminishing marginal utility in terms of more basic facts, I think, is that twentieth century economics has been dominated by narrowly-focused behaviourist and operationalist prejudices which have manifested themselves in a general reluctance to look beyond observed market behaviour for underlying explanations; in particular, to seriously raise questions about the nature of individual welfare, about the possible function of preference strengths in our psychologies and so on. Such prejudices received perhaps their clearest expression in Samuelson's doctrine of revealed preference; and, with specific regard to the concept of diminishing marginal utility, in the Hicks-Allen proposal to do away with the concept in favor of a principle of "increasing marginal rates of substitution".[18] Renaming happy ac-

---

[17]See Edgeworth, Mathematical Psychics, pp. 61-3; Pareto, Manual of Political Economy, pp. 192-3; and Hare, "Justice and Equality" in John Arthur and William H. Shaw, eds., Justice and Economic Distribution (Englewood Cliffs: Prentice-Hall, Inc., 1978), p. 125.

[18]See note 26 of the previous chapter.

cidents does not make them go away, however, and the first responsibility of the theorist must remain to search for explanations of pervasive regularities in nature. Diminishing marginal utility by whatever name is a pervasive regularity, and the fact that it is predicted by the general view of welfare I have been defending, in the right sort of way, provides strong circumstantial evidence for that view.

3. Indirect Interpersonal Comparisons

If the discussion of the preceding sections has been substantially on track, then the kind of interpersonal utility comparisons that we can and do perform are <u>indirect</u> comparisons -- "indirect" because the strengths of preferences are compared via their proportionate weights within individuals' overall systems of preference. To assert that one person's preference for <u>x</u> over <u>y</u> is stronger than another's preference for <u>z</u> over <u>w</u> is to assert that securing <u>x</u> over <u>y</u> means more to the former -- it occupies a greater space on her own personal scale of value -- than securing <u>z</u> over <u>w</u> means to the latter. Similarly, to judge that one person is better off than another in the circumstances is to judge that the former's satisfied preferences comprise a greater proportion of her total possible welfare than do those of the latter. Formally, given utility functions <u>U</u> and <u>V</u> defined over possible worlds for indiv-

iduals A and B respectively, and most and least preferred worlds m,l and o,n in the domains of U and V, A's preference for x over y is stronger than B's preference for z over w if and only if:

$$(3) \qquad \frac{U(x) - U(y)}{U(m) - U(l)} > \frac{V(z) - V(w)}{V(o) - V(n)}$$

and A's level of utility at x is greater than B's level of utility at y if and only if:

$$(4) \qquad \frac{U(x) - U(l)}{U(m) - U(l)} > \frac{V(y) - V(n)}{V(o) - V(n).}$$

(Interpersonal comparisons of preference strength are thus equivalent to comparisons of differences in welfare levels at the appropriate outcomes, since $U(x)-U(y)/U(m)-U(l) = [U(x)-U(l)/U(m)-U(l)] - [U(y)-U(l)/U(m)-U(l)]$.)

Notice that no observations of behaviour or physiology are required to arrive at indirect interpersonal comparisons, beyond those necessary to establish personal interval rankings. In practice, of course, there is no possibility of relying on the vN.M. procedure or anything like it in constructing personal utility functions. What we do instead is estimate the relative importance which individuals attach to outcomes on the basis of various aspects of their behaviour, without attempting any very fine discriminations either in the outcomes or the behaviour. Little and Harsanyi were in a sense right in contending that as a matter of course we perform interpersonal comparisons primarily on the basis of differences in individuals'

expressive behaviour. They mistook the import of that behavioural evidence, however. Expressive behaviour is not relied on to fill the gap between personal interval functions and a full specification of the absolute strengths of individuals' preferences; but rather as a means of estimating the interval scales themselves, given the practical impossibility of applying more rigorous methods. Granted, the estimates are somewhat vague; but for many purposes they are enough.[19]

Indirect interpersonal comparisons are perfectly objective. Assuming that there is a fact of the matter regarding personal interval rankings, and that we can roughly identify most and least preferred worlds for each individual, there is a fact of the matter regarding the proportionate weight of particular preferences within individuals' systems of preference; and an indirect utility comparison is a straightforward comparison of such proportionate weights. It must be acknowledged however that the use of indirect comparisons in formulating social policy is not so straightforwardly an objective matter. Even if it

---

[19]Notice also that it will be no argument against taking individuals' expressive behaviour as an indication of their level of welfare that some people may be quicker or more vocal in expressing satisfaction or dissatisfaction, and that we have no independent means of determining this. In estimating the relative importance that an individual attaches to an outcome, what is important is her behaviour vis-à-vis that outcome in relation to her behaviour vis-à-vis other outcomes; and that is a matter for more-or-less direct observation.

is a plain matter of fact that the outcome of some policy means just as much in proportionate terms to you as the outcome of another policy means to me, it is no simple matter of fact that if someone has to choose between the two policies, other things equal she should flip a coin. But this is only to be expected. Even if preferences did have absolute strengths and it was possible to compare them across individuals, the direct use of such comparisons in making policy recommendations would involve the same minimal normative commitment, which really amounts only to the proviso that policy judgments be based exclusively on the relevant facts of the case. So far as the objectivity of indirect utility comparisons is concerned, the important point to notice is that we now have in hand a reply to Robbins' Brahmin.[20] The Brahmin, recall, maintained that he was ten times as capable of happiness as "that untouchable over there". He was simply wrong about that; given the relational nature of preference intensity, there is no objective sense in which his overall capacity for satisfaction could be ten times that of the untouchable. Of course, it remains open to the Brahmin to attempt to mount an argument to the effect that whether or not he is ten times as capable of happiness as others, his own welfare should be counted for ten times as much in formulating social policy. But that avenue would be available in any

---

[20]See p. 36 above.

case. I presume that whatever argument he produces is not likely to prove very convincing.

On a related note, within the literature on welfare economics one sometimes finds references to individuals' "normalized" utility functions. Normalized utility functions are interval functions which are presumed to have been commensurated in some fashion, so that points on the individual scales are directly comparable. Often the normalization is assumed to have been performed on the basis of having discovered an objective unit of comparison. Some theorists have maintained, however, that regardless of whether individuals may differ in their capacity for satisfaction, it is appropriate for purposes of social choice to normalize individuals' utility functions by equating the values of their maximally satisfying and maximally frustrating outcomes. Frederic Schick has defended this position, e.g.[21] David Gauthier has also suggested that we might assign values of 1 and 0 to each individual's most and least preferred of all logically possible outcomes as a means of performing interpersonal comparisons, though without sanctioning the use of such comparisons.[22]

---

[21] "Beyond Utilitarianism", Journal of Philosophy Vol. LXVII, No. 20 (1971), pp. 657-66.

[22] Morals By Agreement (Oxford University Press, 1986), pp. 240-1. Gauthier proposes the method simply to make the idea of a sum of utilities more concrete, so that he can go on to criticize Harsanyi's derivation of average utilitarianism.

Alert readers will have noticed that comparing the strengths of individuals' preferences in accordance with formula (3) above is formally equivalent to normalizing their utility functions with respect to most and least preferred worlds, and then comparing simple differences between the utility indices of outcomes, rather than comparing ratios of differences. Likewise, formula (4) tells us that welfare levels can be compared across individuals whose utility functions have been normalized in this fashion by directly comparing the utility indices of outcomes. In view of this fact, it might be supposed that indirect interpersonal comparisons really depend on adopting a normative stance towards interpersonal comparisons similar to the one defended by Schick. But that would be a mistake. The formal equivalence of Schick's procedure and ours masks fundamental differences in their underlying motivation.

Given that indirect comparisons are invariant with respect to positive linear transformations, there is of course nothing to prevent us from adopting normalized scales and then proceeding to directly compare utility indices and utility differences if we wish to so; our discussion has shown that in this sense, normalization with respect to most and least preferred worlds is quite innocent. But we <u>needn't</u> bother with normalizations -- indirect comparisons can proceed perfectly well without

scale recalibrations of any sort -- and there is another sense in which the method of normalization is not entirely innocent, since it obscures the fact that the real comparisons are taking place indirectly, via the proportionate weight which preferences carry within individuals' overall systems of value.

Indirect utility comparisons might be viewed with some justification as solving the units of comparison problem by taking the appropriate units of comparison to be individuals themselves; or rather, by taking the appropriate units to be individuals' total systems of preference. The rationale for doing so, however, is not that individuals should count for the same overall in matters of social policy. It rather stems from our understanding of preference intensity, from the fact that it is only within the context of someone's total system of preferences that any preference can <u>have</u> a determinate strength. To suppose that indirect comparisons implicitly depend on a kind of normative presupposition in favor of treating persons equally seems to me to get the matter exactly backwards. Indeed, it is not wholly implausible to suppose that the intuitive force of the idea that individuals should be treated with equal concern and respect depends, at least to some extent, on our implicit recognition of the natural integrity of their systems of value; on recognizing that apart from these systems, there cannot be degrees of

preference and welfare at all.

Finally, a word on the relation of indirect utility comparisons to the so-called "ordinal" comparisons which have recently received attention in the social choice literature. An ordinal ccmparison is supposed to be a comparison of levels of welfare which does not imply anything regarding the extent to which one person is better off than another. Notice that although formula (4) above is presented as a simple inequality, the proportionate representations of levels of welfare on the two sides of the inequality provide the theoretical basis for judgments concerning degrees of difference in the welfare levels of different individuals. I.e., the proportionate construal will in principle allow us to say not only that A would enjoy a higher level of welfare at x than B would at y; but that A would enjoy, say, a 10% higher level of welfare at x than B would at y. An ordinal intercomparison, on the other hand, is supposed to tell us only who is better off than whom in the circumstances, not by how much.

The reason that ordinal comparisons have recently attracted attention is that if one presupposes the possibility of performing such comparisons, it can be shown that there exist social welfare functions which will satisfy at least a minimal set of intuitively appealing conditions on social choice. One way of interpreting Arrow's celebrated General Possibility Theorem is as

showing that there is no general way of combining individual preferences into a plausible social ordering of outcomes in the absence of interpersonal utility comparisons of some sort. In originally demonstrating the theorem Arrow defended a general scepticism regarding cardinal utilities and interpersonal comparisons,[23] and in addition to imposing minimal conditions on the social welfare function such as weak Paretianism, imposed constraints which insured that it would be insensitive to utility comparisons of any sort, even if they were in principle possible. The result is well known: There are no social welfare functions which satisfy all of Arrow's original conditions. If we relax the conditions slightly in order to allow for the possibility of interpersonal comparisons, however, it can be shown that there are in fact methods of aggregating individual preferences which meet the revised set of conditions.

In particular, if the original conditions are weakened just enough to allow for ordinal utility comparisons, there are exactly two principles which meet the conditions: "lexical maximim" and "lexical maximax". The former is a lexical version of Rawls' difference principle, except that it operates directly in terms of utilities rather than in terms of individuals' expectations for primary goods; it states that one outcome is socially better than another if

---

[23] Social Choice and Individual Values, pp. 9-11.

the worst off individual in the former enjoys a higher level of welfare than the worst off individual in the latter, that if the worst off individuals enjoy equal levels of welfare then an outcome is socially better than another if the next worst off individual in the former is better off than the next worst off individual in the latter, and so on. Lexical maximax is an inversion of the former principle which states that rather than lexically maximizing the welfare levels of the worst off persons in society, we should maximize the welfare levels of the best off.[24]

---

[24]For a clear discussion see Arrow, "Extended Sympathy and the Possibility of Social Choice", American Economic Review Papers and Proceedings 67 (1977), pp. 219-25; reprinted in Collected Papers Volume 1: Social Choice and Justice, pp. 147-61. The original result is due to Steven Strasnick, "Social Choice and the Derivation of Rawls' Difference Principle", Journal of Philosophy 73 (1976), pp. 85-99; and to Peter J. Hammond, "Equity, Arrow's Conditions, and Rawls' Difference Principle", Econometrica 44 (1976), pp. 793-804. See also C. d'Aspremont and L. Gevers, "Equity and the Informational Basis of Collective Choice", Review of Economic Studies 44 (1977), pp. 199-210; and Sen "Welfare Inequalities and Rawlsian Axiomatics" in Robert E. Butts and Jaakko Hintikka, eds., Foundational Problems in the Special Sciences (Dordrecht: D. Reidel, 1977), pp. 271-92.

Technically, the result is derived by assuming the existence of a real-valued utility function for each individual, and then imposing an "invariance condition" on the social welfare function which states that the social ordering of outcomes must not vary with certain admissible transformations of the individual utility functions. Ordinal invariance requires that the social ordering not vary with any positive monotone (i.e., order-preserving) transformation of anyone's utility function; this is equivalent to the supposition that no interpersonal comparisons of any sort are possible, and Arrow's original impossibility result follows immediately. Co-ordinal invariance, on the other hand, requires only that the

Interest in ordinal comparisons has no doubt been bolstered by the fact that the formal result of allowing such comparisons in social choice theory is strikingly in tune with Rawls' ethical position. Some theorists appear to be of the view, however, that ordinal comparisons are in themselves more respectable than utility comparisons of other sorts, quite apart from their specific implications for ethical theorizing.[25] This is, I think, an error. The view that ordinal comparisons are more respectable than other sorts of comparisons presumably stems from the supposition that, as in the case of individual utilities, ordinal judgments are in principle less demanding than cardinal ones, in the sense that they require less information about what things individuals value and how they value them. This is no doubt true in the case of individual utilities. But it is not true (or at any rate, not

---

social ordering not vary when everyone's utility function is subjected to the same order-preserving transformation. This has the effect of allowing for ordinal intercomparisons, since if my original utility function assigns a higher value to some outcome than yours does, any simultaneous monotone transformation of the two functions will preserve this result. Informally, the reason why co-ordinal invariance rules out all social decision rules except maximin and maximax is that allowing any and all simultaneous monotone transformations of individuals' utility functions makes it impossible for the social welfare function to be sensitive to information about how much better off some people are than others, with the result that the only welfare levels it is possible to fix on are those of the worst off and best off individuals. A weak "equity" condition can be imposed to rule out focusing on the best off, leaving lexical maximin as the only rule satisfying all of the conditions.

[25] Cf. for example Arrow, ibid., p. 151.

true to anything like the same extent) in the case of interpersonal utilities. We noticed in the last section of the previous chapter that value is not a simple relation holding between an individual and some object or event or state of affairs; properly speaking we only value things (and disvalue them) in relation to other things. Ordinal judgments of personal utility are quite consistent with this fact: we judge simply that A prefers x to y, without attempting to say anything about how much she prefers the former to the latter.

It is not at all clear, on the other hand, that ordinal judgments of interpersonal welfare (insofar as they might be presumed to be more respectable than other sorts of interpersonal judgments) are consistent with the three-part nature of the value relation. The idea underlying ordinal comparisons is that it is possible for us to say that A values an outcome x more than B values y, but not by how much. Given the nature of value, however, we are entitled to ask: A values x *in relation to what* more than B values y (in relation to what)? In terms of preferences: Obviously A cannot prefer x more than B prefers y *simpliciter*; that is flatly incoherent, since no-one can prefer an outcome *simpliciter*. The point here is that if it makes any sense at all to suppose that a person values some outcome more than another values some other outcome, this must ultimately be understood somehow in terms of the

subjective importance which they attach to some outcomes in relation to others; comparisons of welfare levels, whether ordinal or not, depend on judgments concerning the preference strengths of different individuals.[26]

The present point may be confirmed by focusing on the supposed "operational content" of an ordinal utility comparison. Most theorists who discuss such comparisons rely on the idea that we perform them via the method of "extended sympathy" -- roughly, the method of putting oneself in another's shoes (understood to include putting on the other's preferences and leaving one's own behind), and then attempting to decide whether we would rather be one person or another under various outcomes.[27]

---

[26]Note that an ordinal comparison of welfare levels may still be less demanding than a non-ordinal one, in the sense that a rough comparison of the subjective importance which individuals attach to particular outcomes in relation to their maximally satisfying and maximally frustrating outcomes might unequivocally support the claim that one person would enjoy a higher level of welfare than another at a given outcome, but not a more detailed claim concerning how much higher the former's level of welfare would be. The crucial point is that ordinal intercomparisons depend on comparing individual preference strengths, whether the latter comparisons happen to be rough or more precise, and hence that it is a mistake to think that ordinal comparisons are theoretically more respectable than comparisons of other sorts. Ordinalism regarding interpersonal comparisons cannot receive any direct support from ordinalism with respect to individual utilities.

[27]Sen distinguishes the method of 'introspective welfare comparisons' from the method of 'introspective as if choice'. The former involves putting to oneself the question "Do I feel I would be better off as person i in social state x rather than as person j in social state y?", whereas the latter involves asking not "In which position do I feel I would be better off?" but rather "Which position would I choose?". ("Interpersonal Comparisons of

I do not doubt that we often do make rough comparisons of individuals' levels of welfare via sympathetically identifying with them, and that such comparisons are (or at least may be) as objective as any we are normally in a position to make. It would be a mistake, however, to suppose that the fact that interpersonal comparisons based on the method of extended sympathy focus explicitly on only a single outcome for each individual supports the view that such comparisons are in principle less demanding than comparisons of preference strength. Suppose that I judge that all things considered I would rather be myself in the circumstances than some other person. According to the method of extended sympathy, I arrive at this judgment by sympathetically identifying with this other person, including identifying with her subjective estimation of which things are important and which aren't, and then deciding

---

Welfare" in M. Boskin, ed., Economics and Human Welfare: Essays in Honor of Tibor Scitovsky, New York: Academic Press, 1979; reprinted in Amartya Sen, Choice, Welfare, and Measurement, Cambridge, Mass.: MIT Press, 1982, pp. 264-81.) Although the difference between the two methods appears to be slight, the former may be theoretically preferable for reasons originally raised by Rothenberg in relation to Harsanyi's use of something like the method of extended sympathy in his derivation of average utilitarianism (The Measurement of Social Welfare, pp. 268-9), and more recently by David Gauthier in the same vein ("On the Refutation of Utilitarianism" in Miller and Williams, The Limits of Utilitarianism, pp. 144-63). Curiously, Sen cites both methods as counterexamples to Robbins' claim that judgments of interpersonal welfare are normative rather than descriptive, on grounds that in both cases one is describing aspects of one's own psychology (in the former, what one feels, in the latter what one would choose). I doubt that Robbins would have been impressed.

that I would rather be in my own shoes than in hers. But how is it that I manage to sympathetically identify with her personal estimation of the importance of various things? I can of course get a rough idea of the relative importance which she attaches to various outcomes on the basis of her behaviour, especially her expressive behaviour. But how do I translate this relative importance into something which has interpersonal significance? How do I know, e.g., that she isn't a thousand times better off with even her worst of all possible outcomes than I am with my best of all possible outcomes?

The answer, I think, is that I obviously can't know this on the basis of anything that I can observe about her behaviour. (At any rate, if I can know this, there is no problem of interpersonal comparisons, and there never was.) The method of extended sympathy does offer a practical guide for formulating rough judgments of interpersonal welfare; but only because it relies on assigning a roughly equal weight to individuals' total possible welfares, thus allowing us to deduce the levels of welfare which they would enjoy at given outcomes on the basis of the importance which they assign to those outcomes in relation to their maximally satisfying and maximally frustrating outcomes. The method of extended sympathy, in other words, is precisely a method of performing indirect welfare comparisons; it is not an alternative to performing such

comparisons. And inasmuch as indirect interpersonal comparisons will in principle provide us with information concerning degrees of difference in individuals' levels of welfare, there can be no theoretical basis for refusing to avail ourselves of this information in social choice.

CHAPTER 4: UTILITY AND EQUALITY

1. Taking Seriously the Distinction Between Persons

Utilitarianism has come in for a enormous amount of criticism in recent decades. It is sometimes hard to escape the conviction, when reviewing the recent literature, that Williams' prediction concerning the fate of the theory[1] has not been borne out in the main because critics delight in devising new objections to it, or new twists on old objections. My goal in these final two chapters is to indicate some of the ways in which many of the objections and twists issue from confusions regarding the form and content of utilitarian theorizing.

In one way or another, concerns about the distributive consequences of utilitarianism rank high on virtually every contemporary critic's list of complaints. We are often reminded that under suitable (perhaps improbable) circumstances, acting on the advice of an unrestrained principle of utility could lead to radically inegalitarian and intuitively unacceptable distributions of goods. Some have gone further, insisting that the principle of utility sanctions unacceptably inegalitarian outcomes in quite normal circumstances. Rawls among others has attempted to

---

[1] "The important issues that utilitarianism raises should be discussed in contexts more rewarding than that of utilitarianism itself. The day cannot be too far off in which we hear no more of it." ("A Critique of Utilitarianism", p. 150)

trace these supposed faults to certain deep theoretical shortcomings: Utilitarianism does not take seriously the distinction between persons, in his view, and hence is in principle incapable of attending to issues relating to the distribution of goods among them.

Objections having to do with the inegalitarian tendencies of utilitarianism are by now so commonplace that it is perhaps easy to overlook their slightly paradoxical air. Harsanyi has noted that until comparatively recently, utilitarianism was thought to be a highly egalitarian doctrine both in spirit and practice (thought by some, indeed, to be altogether too egalitarian in its distributive consequences).[2] Common sense appears to back this earlier assessment (which is precisely what lends the contemporary view its air of paradox). If we step back from theory for a moment and focus on our ordinary judgments of interpersonal welfare as a base-line for assessing the distributive implications of utilitarianism, I think it will be generally acknowledged that the principle of utility is likely to recommend substantial equality in the

---

[2] "Nonlinear Social Welfare Functions: Do Welfare Economists Have a Special Exemption from Bayesian Rationality", p. 68; cf. Arrow, "Formal Theories of Social Welfare", p. 121. Harsanyi specifically cites Robbins as thinking that utilitarianism might turn out to be too egalitarian. Among actual utilitarians, Edgeworth seems to have been most exercised by the egalitarian character of the theory; see Mathematical Psychics, pp. 77-80, as well as "The Pure Theory of Taxation" in Papers Relating to Political Economy, Vol. II (London: Macmillan and Co., 1925), pp. 63-125.

distribution of goods -- more equality than exists in most societies today, in all likelihood, since barring implausibly strong assumptions about incentive effects and so on, our common sense judgments would seem to indicate that aggregate welfare could be increased substantially via redistributions of income and wealth in excess of those currently fashionable.

My main contention in what follows is that common sense is fully supported by theory in this regard. Distributive objections to utilitarianism are not content-neutral as regards the nature of individual welfare and interpersonal utility comparisons. More precisely, to the extent that objections concerning the supposedly inegalitarian tendencies of utilitarianism are formulable at all in the absence of concrete assumptions about welfare and interpersonal comparisons, their force <u>as</u> objections depends entirely on presuppositions which cannot be maintained if the argument of the preceding chapters is correct. At any rate, that is the conclusion I shall urge in the remainder of this chapter. The present section is devoted to showing that Rawls' formal objections to utilitarianism cannot be sustained against a utilitarianism based on indirect utility comparisons. Subsequent sections attempt to assess the weight of more substantive distributional complaints in light of the account of welfare and interpersonal comparisons defended in Chapter 3. In the

final section we briefly turn our attention to more practical concerns; there I argue that the concept of needs as it plays a role in the formulation and implementation of social policy is underwritten by the principle of utility, in conjunction with our practical inability to perform detailed comparisons of individuals' welfares.

We begin, then, with Rawls' charge that utilitarianism fails to take seriously the distinction between persons. The charge stems from a particular understanding of how the utility calculus operates, which is in turn supported by a particular view of the motivations underlying the utilitarian position. The "natural course of reflection" by which one may arrive at utilitarianism, according to Rawls, involves extending the principles of rational choice appropriate for a single individual to society as a whole. Just as the rational individual naturally balances personal gains and losses against each other in determining how best to promote her own welfare, so too the utilitarian thinks (at least, so Rawls thinks the utilitarian thinks) that the rational way to proceed at the societal level is to balance gains for some individuals against losses for others in determining how best to promote aggregate welfare. This balancing act is supposed to occur via the imaginative feats of an ideally rational and impartial sympathetic spectator, who identifies with and experiences the desires of each individual as though they were her own. In this

fashion the sympathetic spectator manages to gauge the intensity of each individual's desires and assign them their proper weight within a single coherent system of desire, which may then serve as the basis for rational social choice. But in thus "conflating all persons into one," Rawls maintains, utilitarianism fails to properly acknowledge their "plurality and distinctness". This "conflation of all desires into one system of desire" thoroughly undermines whatever claim utilitarianism might have to be an adequate moral theory.[3]

This criticism is in some danger of becoming a standard weapon in the critic's arsenal, and it is thus worth investing a bit of effort in carefully defusing it.[4] Utilitarianism certainly does call for balancing the gains of some individuals against the losses of others in assessing social policy. (More accurately: Utilitarianism will call for balancing gains to some against losses for

---

[3] A Theory of Justice, pp. 22-29 and 187-88. I presume that this is one of the "more important objections to utilitarianism" which Rawls refers to in the passage quoted in Chapter 1 above.

[4] H.L.A. Hart refers to it as "the distinctively modern criticism of utilitarianism"; see "Between Utility and Rights", Columbia Law Review 79 (1980), p. 829. Cf. also David A.J. Richards, "Prescriptivism, Constructivism, and Rights" in Hare and Critics: Essays on Moral Thinking, p. 118: "The kinds of objections which Hare wishes to rebut are now familiarly explained in terms of a central blunder of utilitarian moral reasoning, namely its failure to take seriously the separateness of persons." Richards goes on to develop the criticism in exactly the way that Rawls does, and then attempts to wield it specifically against Hare's position. For the latter's response, see "Comments on Richards", ibid., pp. 255-60.

others in all realistic circumstances -- i.e. circumstances in which there is no outcome which is Pareto superior to all others.) It is far from clear, however, that the theory warrants such balancing by analogy with the practical deliberations of individual persons. Whether or not it is true that some utilitarians may have talked themselves into their position via a simple extension of the principles of rational choice for individuals to the societal level, few if any have attempted to justify the principle of utility on that basis. Far simpler, it seems to me, is the hypothesis that utilitarianism warrants the balancing of gains and losses via the observation that the only way to completely avoid such balancing, in realistic circumstances, is by either insisting on something like a principle of strict equality of distribution (of either goods or welfare), or else ignoring individuals' welfares altogether. Whether social policy should be assessed exclusively in terms of individual gains and losses is of course a further question, precisely the one which most of the present chapter is devoted to answering. My point is simply that we don't need to appeal to hidden motivations or justifications for the view that judging the worth of social policies depends to some extent on balancing gains and losses; a fully adequate justification resides in the unpalatableness of the alternatives. Notice in this regard that the difference principle itself warrants balancing

losses to some against gains for others; as numerous authors have complained, the principle grants infinite weight to gains for the least advantaged members of society, in comparison with the losses of those more advantaged. To be sure, no actual totting up of gains and losses occurs when applying the difference principle; but that is only because it has been pre-ordained that the total in the loss column, whatever it is, will be insufficiently large.[5]

Note further that the fact that utilitarianism calls for balancing gains against losses in assessing social policy in no way implies a conflation of all desires into a single system of desire, any more than the difference

---

[5]Some may find it odd to speak of the losses which the more advantaged would suffer at the hands of the difference principle, but this is nevertheless the correct formulation. The outcomes of alternative social policies are so many different world-lines stretching into the future, and hence the gains and losses at issue are properly measured in terms of the goods or welfare which would accrue to individuals under one possible future, in comparison with what they would receive under other possible futures. We sometimes have a tendency to think of gains and losses in terms of the amount of material possessions or welfare that someone will have at some time in the future as compared with what she has now: two yo-yos today, three tomorrow, hence a net gain of one yo-yo; 1,000 "utiles" today, 500 tomorrow, a net loss of 500. On this construal, at least some of the so-called losses of the more advantaged are not really losses at all, but rather represent possible gains foregone. This way of construing gains and losses would beg all sorts of questions in the theory of social choice, however, since any attempt to balance gains and losses measured in this fashion would implicitly sanction present distributions. (That is not to say, of course, that present distributions can in principle have no bearing on the assessment of policy; only that if they are to have a bearing, this needs to be argued for.)

principle implies such a conflation. (It might be thought that a conflation is implied by the fact that the principle of utility operates directly on gains and losses in welfare, whereas Rawls' version of the difference principle operates in terms of primary goods. It is difficult to see why this should make a difference, however. Indeed, one might think that it is the doctrine of primary goods, together with such devices as representative least advantaged persons, which implies an illegitimate conflation of individual desires. After all, different people in fact value different primary goods differently. The doctrine of primary goods is precisely an attempt to average over such differences and so arrive at a kind of universal "one size fits all" utility function. If this does not count as a failure to acknowledge individuals' plurality and distinctness, what does?)[6]

---

[6]Cf. Hare, ibid., p. 258: "It is Rawls who fails to take [persons] seriously, by forcing on them a list of 'primary goods' or 'common goods', culled from his own intuitions -- goods which some of them may not in all circumstances prize as much as he does." Rawls attempted to pre-empt criticisms of this sort by noting that primary goods are in no way forced on people, since they are entirely free to reject the goods allotted to them by the difference principle after the veil of ignorance is lifted. But the pre-empt is not entirely satisfying, inasmuch as it does nothing to address the concerns of POPs who recognize that their actual life plans may require little in the way of the primary goods on Rawls' list, and who wonder why the basic structure of society should be so overwhelmingly biased in favor of promoting the life plans of those who do place great stock in the goods on the list. Lest it be thought that this is an idle concern, since society is not anyway in a position to promote life plans which don't involve much in the way of primary social goods, notice

It is perhaps for this reason -- to fill the gap between balancing gains and losses and the hypothesized conflation of desires -- that the notion of an ideal sympathetic spectator is brought into the discussion. One is inclined to point out at this stage, what in any case Rawls must be well aware of, that ideal spectators have played far less of a role in utilitarian theorizing than his reconstruction suggests. More importantly, to the extent that something like an ideal spectator does play some role in the theories of some utilitarians, we haven't as yet been given any indication of why this should lead them to ignore the boundaries between persons. Hare has perhaps come closer than any other utilitarian to granting a role to a sympathetic spectator in his theory, due to the particular way in which he attempts to derive utilitarianism from his metaethical position. Hence if anyone is guilty of making the mistake which Rawls attributes to utilitarians generally, we might expect that it would be

---

that in point of fact life plans which don't centrally involve the goods on Rawls' list often conflict with those that do. Consider someone who values spending most of her days wandering through pristine wilderness and communing with nature. Her plans may clearly be jeopardized to some extent by institutions directed towards increasing the supply of income and wealth among other primary goods. Perhaps, given others' values and life plans, justice will in the end require basic institutions whose effect will be to constrain the life plans of nature lovers; but it seems more than a little strange that this result might be secured a priori, through ignoring the interests of nature lovers altogether in designing the basic structure of society.

him. But I can find no evidence in Hare's work, early or late, which suggests that he or his theory conflates either persons or their desires, and he has recently disavowed any such conflation.[7]

What exactly is it about the notion of an ideal spectator which prompts critics like Rawls to suppose that any theorist who appealed to such a notion would be guilty of conflating persons and their desires? Presumably it is not the mere fact that the ideal spectator is supposed to have perfect powers of sympathetic identification. The ability of someone to experience my desires as though they were her own, and your desires as though they were her own, and similarly for each other person's desires in turn, does not seem to require or imply a conflation of all of our

---

[7]In responding to Richards' attack Hare writes (ibid., pp. 256-7):

> Would the people who repeat this criticism of Rawls say that an impartial arbitrator is failing to take seriously the difference between persons if he administers even-handed justice by treating the equal interests of two different parties as ceteris paribus of equal weight, just as in our prudential judgments we so treat our own interests? Does such an arbitrator not know that he is dealing with two different persons (can he not count)? Rather, he is, like the utilitarian, trying to do justice between the parties, showing them equal concern and respect.

And immediately following:

> Richards is correct in finding this affinity between ideal-observer theories and my own utilitarianism (see MT 44 and refs.). Since God is the paradigm ideal observer, the 'blunder' is committed by most forms of Christian ethics. God too knows that we are different people, but he loves us all equally.

desires into a single system. It is difficult to be sure what the critics have in mind here, but it may be that they arrive at their conclusion not by supposing merely that an ideal spectator would experience each person's desires as if they were her own; but rather by supposing that she would experience everyone's desires, all at once, as if they were her own. In other words, the sympathetic spectator is not simply a master at putting herself in others' shoes; she is actually a master contortionist who is able to wear everyone's galoshes at the same time. Notice that this interpretation of the sympathetic spectator seems to be required if we are to take seriously the claim that utilitarians arrive at their position via a literal extension of the principles of rational choice for individuals to the societal level; otherwise the supposed extension is not much more than a loose analogy, reflecting the unexceptional fact that rational policy makers will appropriately weigh the pros and cons of alternative courses of action, just as rational individuals do.[8]

If this is what the critics have in mind, then it is easy to see why they think that the notion of an ideal

---

[8]Cf. Richards, ibid., pp. 118-19: "The approbations of the ideal observer would, in short, reflect the utilitarian principle in precisely the way that literally conflates the rationality of prudence with the reasonableness of ethics: the approbations of the ideal observer, expressive of the greatest pleasure over pain in himself, literally reflect (via sympathetic identification) the aggregate of pleasure over pain in the moral community at large." (my emphasis)

spectator essentially involves a conflation of individuals' systems of desires. But then it is even more of a mystery why they should be inclined to think that utilitarians could have blundered so badly. For my own part, I do not think that the notion of a sympathetic spectator as the critics apparently want or need to construe it is even coherent. I cannot see how such a spectator is going to remain ideally rational and impartial and prefer x to y and y to x all at the same time; far from resulting in a single coherent system of desire, the attempt is more likely than anything else to land her in an ideal mental institution. In any case, regardless of whether the notion is coherent, I'm quite sure that utilitarians don't typically rely on it, either explicitly or implicitly. Hare certainly doesn't.

To summarize the discussion of the last few paragraphs: (1) There is scant evidence to support the contention that utilitarians arrive at their position by extending the principles of rational choice for the individual to the societal level, or by placing considerable weight on the notion of an ideal spectator, or some combination of the two;[9] and (2) even if they did, there is no evidence to

---

[9]Rawls himself seems to have realized that his gloss on "the most natural way" of arriving at utilitarianism strains the principles of sound philosophic interpretation. In n.10 on p. 24 of A Theory of Justice he directs us to a long note which appears on p. 188 for "references to utilitarians who explicitly affirm [the] extension" of principles of individual rationality to society as a whole.

support the claim that this would lead them to conflate persons and their desires. None of this actually shows that utilitarianism does take seriously the distinction between persons, however. Forgetting about individual rationality and ideal spectators &c. for the moment, isn't it possible that the utility calculus does as a matter of fact conflate persons and their desires in ranking social policies?

In a word, no. It is important to recognize, first of all, that no utilitarian or anyone else, regardless of what the motivation was, could literally combine everyone's desires into a single coherent system of desire as Rawls suggests. The reason is simply the one given two paragraphs previous: there is no such thing as a coherent system of desire in which x is preferred to y and y to x at

---

There he cites passages from C.I. Lewis, Smart, and Hare which appear to support his reconstruction. The first can hardly be classed as a luminary in the history of utilitarian theorizing, however, while the latter two are avowed non-cognitivists who explicitly defend their positions in ways quite different from what Rawls suggests is the normal course of reflection (Smart by specifically addressing himself only to those who share his "generalized benevolence"; Hare by imposing the "logic" of moral discourse on individual prescriptions.) Rawls continues: "Among the classical writers the conflation of desires into one system is not to my knowledge clearly asserted." He then goes on to quote passages from Edgeworth and Sidgwick which he thinks hint at the motivation he attributes to utilitarians generally. But the passage from Edgeworth (not to mention numerous other passages in Mathematical Psychics) explicitly defends utilitarianism (or rather, "social mechanics") by analogy with celestial mechanics, not by analogy with individual rationality; while the quotations from Sidgwick appear to support Rawls' interpretation only in the sense that they do not contradict it.

the same time.[10] Thus if the utility calculus does conflate persons and their desires, it must accomplish this in some way other than by combining their desires into a single system. I can think of one other way in which the utility calculus might be supposed to result in a conflation of the requisite sort, and I suspect that at base it must be this that the critics have in the back of their minds when they advance their objections. I refer to this conception of how the utility calculus operates as the "stew-pot" interpretation of utilitarianism. On the stew-pot interpretation, the utilitarian chef begins by throwing everyone's preferences into a big pot, discarding the empty shells.[11] She then stirs the preferences all together, and proceeds to figure out how to satisfy the most preferences in the pot, allowances made for different intensities. In

---

[10]Perhaps the one coherent system at issue is supposed to be just the social welfare function which results from aggregating individual utility functions in accordance with the principle of utility, so that in a two-person society if A prefers x to y just as much as B prefers y to x, then x and y will be indifferent in the one coherent system of desire. If this is all that Rawls had in mind, it's hardly the basis for an objection.

[11]Cf. Richards, ibid., p. 121: "In a real sense, [Hare's] argument does not take persons or [the] personal point of view seriously, for the view gives weight to other persons not as persons but, externally and impersonally, as containers of preferences with which the prescriber projectively identifies." Hare responds: "The rhetorical expression 'containers of preferences' presumably means, when put in ordinary English, 'people who have preferences, i.e. prefer one thing to another'. Does Richards think that morally irrelevant? If Tom cuts Jane's finger off, is it not a moral consideration that she would prefer to keep her finger? Or is she just a 'container of preferences'?" (ibid., p. 258)

this way she manages to conflate everyone's desires not by combining them into a single system of desire, but rather by lumping them all together in a way which makes it impossible to tell which preferences belong to which people, or even how many individuals' preferences were originally placed in the pot. The social rankings which eventually get served up thus do, in a real sense, ignore the boundaries between persons.

Utilitarianism does not work like this. I do not merely mean that the "shells" or "containers" which the critic imagines the utilitarian chef throwing away when she is assembling her ingredients provide essential seasoning, since many of the preferences in the pot happen to involve their shells in an essential way (Hare's point); though that in itself is of course ample reason to look for a more plausible metaphor. My point is stronger: The utility calculus cannot work in this fashion. The stew-pot interpretation rests on the supposition that preferences have their intensities as monadic properties, so that the intensity of each preference will remain intact when we extract it from someone's system of preferences and stir it in with a bunch of other preferences of indeterminate origin. If preference intensities are actually relational in the way described in the previous chapter, however, then whatever else the utility calculus does, it cannot indiscriminately mix together different individuals' preferences

in this fashion and then hope to pick and choose among the tastiest morsels; because by the time all the ingredients are in the pot, the morsels aren't going to have any tastes (strengths) left at all.

The stew-pot interpretation of utilitarianism is thus incoherent, inasmuch as it depends on a mistaken conception of preference intensity.[12] I do not know how many critics have actually had something like the stew-pot interpretation in mind when advancing their objections, though I have failed to discern anything else in their criticisms which would support the claim that utilitarianism conflates persons and their desires. However that may be, the argument of the previous chapter provides us with a principled reason for dismissing the charge, regardless of what the critics have in mind. The fact that a certain prefer-

---

[12]We need to be a bit careful here. What we should properly say is that the stew-pot interpretation is incoherent insofar as it would incline one to think that utilitarianism does not take seriously the distinction between persons. It would be possible to elaborate the stew-pot metaphor in a way which took into account the relational nature of preference intensity -- we need only imagine the utilitarian chef carefully disassembling each person's system of preferences and marking each preference with a proportional strength before tossing it into the pot: this one accounts for 2% of someone's possible welfare, this one for 1%, and so on. But since any such scheme must implicitly recognize the natural integrity of individuals' systems of value, in the sense that a decision to satisfy one of the preferences in the pot at the expense of another would involve ineliminable reference to all of the other preferences of the individuals in question, it couldn't justly be claimed that such a scheme would involve a conflation of persons or their desires in any interesting sense.

ence happens to number among my preferences rather than yours plays an essential role in the utility calculus, since it is only in the context of my complete system of preferences that a given one of them can meaningfully be assigned a determinate and hence interpersonally comparable magnitude at all. Utilitarianism does take seriously the distinction between persons.[13]

2. Utility Monsters and Satisfaction Machines

The fact that utilitarianism is not formally defective

---

[13]Ronald Dworkin is one theorist who has come close to explicitly endorsing something like the stew-pot interpretation of utilitarianism. In responding to an objection of Hart's he writes:

> Someone who reports more preferences to the utilitarian computer does not (except trivially) diminish the impact of other preferences he also reports; he rather increases the role of his preferences overall, in comparison with the role of other people's preferences, in the giant calculation.[14]

Dworkin is clearly presupposing here that preferences bear their intensities as monadic properties. If I acquire new preferences, this does not affect the strengths of my old ones. The new preferences simply get tossed into the pot with all the rest; the weight which my original preferences carry in the utility calculus is not thereby diminished, except in the trivial sense that all of the preferences that were originally in the pot, mine included, account for slightly less of the total.

[14]"Rights as Trumps" in Jeremy Waldron, ed., Theories of Rights (Oxford University Press, 1984), p. 160. Dworkin had argued in Taking Rights Seriously (Cambridge, Mass.: Harvard University Press, 1977, Ch. 12) that the utility calculus would fail to treat people with equal concern and respect unless it disregarded all external or other-regarding preferences. Hart's objection was that this would result in under-counting the interests of those with external preferences, and so fail to treat them in turn with equal concern and respect; see "Between Utility and Rights", pp. 836 ff.

in the way that Rawls and others suggest does not show that the theory isn't substantively defective in its handling of distributional matters. As noted above, the principle of utility does call for balancing gains to some individuals against losses for others, and insists moreover that such balancing is the only thing which is relevant to assessing social policy. This has led countless theorists to suppose that, in principle at least, utilitarianism is compatible with highly inegalitarian and obviously unacceptable distributions of goods and/or welfare. Thus Hart writes:

> [S]ince utilitarianism has no direct or intrinsic concern but only an instrumental concern with the relative levels of total well-being enjoyed by different persons, its form of equal concern and respect for persons embodied in the maxim "everybody to count for one, nobody for more than one" may license the grossest form of inequality in the actual treatment of individuals, if that is required in order to maximize aggregate or average welfare. So long as that condition is satisfied, the situation in which a few enjoy great happiness while many suffer is as good as one in which happiness is more equally distributed.[15]

This passage nicely summarizes, I think, the prevailing philosophical attitude towards the distributive implications of the principle of utility. In order to forestall misunderstandings concerning the arguments to follow, I wish to go on record now as being in strict agreement with the summary, strictly interpreted. It is axiomatic that utilitarianism will rank alternatives which are equivalent

---

[15]Ibid., p. 830.

with respect to promoting aggregate welfare as being socially indifferent, regardless of how that welfare would be "distributed" among individuals, and I shall not attempt to persuade anyone otherwise.[16] I do wish to argue, however, that presuppositions about the nature of welfare and interpersonal comparisons have led critics to badly misconstrue the force of this observation. In other words, while it is certainly true that <u>if</u> an alternative which would result in a few people enjoying high levels of welfare while many suffer is equivalent in respect of aggregate welfare to one on which everyone would enjoy roughly equal levels of welfare, then the principle of utility will rank them evenly, I do not think that this counts as an objection to utilitarianism. And I do not think that the critics would see it as an objection either, were it not for the fact that confusions regarding welfare

---

[16]The scare-quotes on "distributed" are intended to ward off any temptation to slip into thinking of welfare as a kind of good which can be distributed like other goods: a little bit for this person, a little for that one, and so on until the available stock runs out. It is all too easy when discussing distributive matters to implicitly hypostatize welfare in this fashion, forgetting that people enjoy given levels of welfare only in virtue of being located in a particular world to which they assign a certain subjective importance in relation to other worlds. Doing so, however, will almost certainly eventuate in criticisms which at base turn on supposing that the utilitarian presumes welfare to have intrinsic moral worth. So far as the present discussion concerned, it may well be that some of the intuitive force of objections like Hart's stems from hypostatizing welfare in this way, and then thinking to oneself that a more adequate theory than utilitarianism would surely dole out the available stock in a more equitable fashion; on this point, see p. 184 below.

and interpersonal comparisons have prompted them to vastly overestimate the frequency of circumstances in which the antecedent of the conditional may be satisfied.

There are two features of indirect interpersonal comparisons which work in conjunction with each other to insure that utilitarianism is a far more egalitarian doctrine than has commonly been supposed. The first is that, according to our proportionate construal of individual welfare, the marginal utility of any good which is valued to any appreciable extent must eventually diminish for all persons. Those who object to the potentially inegalitarian consequences of utilitarianism typically grant that, to the extent that the principle of diminishing marginal utility does hold, their objections may be somewhat muted.[17] As previously noted, however, diminishing marginal utility has usually been regarded as a more-or-less contingent fact about the way humans value things; hence one presumably need not travel very far in logical space to find a world where the principle fails for some people, and in such a world utilitarianism will recommend highly inegalitarian outcomes. But on the proportionate construal of welfare, diminishing marginal utility is not an accidental feature of the way we value things; it is an inescapable consequence of the relational nature of preference intensity, and one which imposes tight constraints on

---

[17]See e.g. Hart, ibid.

patterns of distribution within actual and possible communities.

The second feature of the present account of welfare and interpersonal comparisons which tends towards substantive equality is the fact that, if we compare welfares indirectly, then each person's total system of preferences will carry exactly the same weight in the utility calculus. We noted above in replying to Robbins' Brahmin that, given the relational nature of preference intensity, it is not possible for people to differ in their overall capacity for satisfaction. In other words, assuming that it makes sense to compare individuals' satisfactions at all, and taking necessity to be the dual of possibility, necessarily people possess the same total capacity for satisfaction. It is this rather striking result in particular which is bound to render utilitarianism a more egalitarian theory of distributive justice than its critics suppose.

There are actually two senses of the phrase 'capacity for satisfaction' which we may distinguish. The fact that utilities must be compared indirectly insures that people cannot differ in their overall capacity for satisfaction, in the sense that their total possible welfares -- the interval between their most and least preferred of all possible worlds -- will be weighted equally in the utility calculus. In some contexts, however, it may be useful to distinguish this sense from another one in which indiv-

iduals can differ in their capacity for satisfaction: namely, in their capacity to be satisfied by certain fixed amounts of resources. For example, if the satisfaction of some of my preferences would require more of a certain resource than is presently available, whereas none of yours do, then there is a sense in which you possess, in situ, a greater capacity for satisfaction than I do. We might refer to this notion generally as individuals' "restricted capacity for satisfaction", with different restrictions on the resources available for satisfying preferences giving rise to particular restricted capacities (e.g., restrictions on the amount of particular resources available, technological restrictions, restrictions due to physical possibility, etc.).[18]

The notion of restricted capacities for satisfaction is a kind of efficiency concept, having to do with how effectively a fixed set or bundle of resources will contribute to someone's welfare. One of the reasons that it is useful to distinguish the concept is that it provides a

---

[18]A formal semantics for the notion of restricted capacities can be constructed in terms of possible worlds: The restricted capacity for satisfaction which an individual possesses for an arbitrary set of worlds may be identified with her level of welfare at the world or worlds in that set which she most prefers. Thus the set of physically possible worlds will give rise to individuals' physically restricted capacities for satisfaction, technologically possible worlds to their technologically restricted capacities, and so on. Someone's total or unrestricted capacity for satisfaction is simply her level of welfare at her most preferred of all possible worlds.

dramatic illustration of the way in which a utilitarianism based on indirect interpersonal comparisons differs in its distributive implications from what many theorists have supposed. Critics have sometimes attempted to wield a kind of efficiency-of-satisfaction concept against utilitarianism. E.g., Hart suggests that any egalitarian implications which the principle of diminishing marginal utility might have for utilitarian distributions will tend to be offset by the "failure of the standard assumption that all individuals are equally good pleasure or satisfaction machines, and derive the same utility from the same amount of wealth."[19]

Hart is right to point out that the standard assumption that people are "equally good pleasure or satisfaction machines" is false. (The phrase smacks a little too much of supposing that the utilitarian presumes welfare to have intrinsic moral worth, and regards persons merely as efficient machines for producing it; but leave that aside. Note also that the assumption is "standard" only in the sense that it is sometimes made for purposes of economic

---

[19]Ibid. Cf. also Milton Friedman's discussion of "enormously more efficient pleasure machines" in "Lerner on the Economics of Control", The Journal of Political Economy 55, pp. 406-16. It is worth noting that Edgeworth's attempt to head off the egalitarian implications of his position rested largely on contending that the rich, because of their superior education and refined tastes (and males, by nature!), were more efficient pleasure machines than the poor, and hence that redistributing wealth from the former to the latter would decrease aggregate welfare; see Mathematical Psychics, pp. 77-80.

theorizing in the absence of concrete data about interpersonal utility comparisons.) He is wrong, however, in thinking that this will tend to result in the principle of utility sanctioning objectionably inegalitarian outcomes. The reason is that, in the context of indirect interpersonal comparisons, a "more efficient satisfaction machine" is simply someone who possesses simpler or more modest (i.e., more easily satisfiable) preferences.

Consider once again the situation where I possess preferences which are unsatisfiable given available resources, whereas you do not. Then if all of the available stock of resources must be given to one of us, and aggregate welfare is the only relevant consideration, clearly you should get the entire stock, since your capacity for satisfaction in the circumstances is greater than mine. Of course, in more realistic settings it will be possible to divide the available resources between us, and hence the principle of utility will recommend that you get the entire stock only if you prefer each increment of each resource more than I do (which is unlikely but not impossible, since nothing which has been said so far rules out my possessing a quite bizarre set of preferences whose satisfaction is largely independent of whatever resources are available.) But the point remains that the principle of utility will tend to place a premium on individuals with modest preferences, at least until such time as diminishing marginal

utility begins to render their remaining, unsatisfied preferences as costly in terms of resource requirements as the preferences of more demanding (greedier?) individuals.

If modest preferences have first claim on available resources, however, then precisely in virtue of the fact that they <u>are</u> modest, their satisfaction will tend to leave more resources available for satisfying the costly preferences of others. Thus from the egalitarian's perspective, the principle of utility may be viewed as a kind of self-correcting principle of distribution. Insofar as some people are more efficient satisfaction machines than others, the principle will favor increasing their complement of resources and hence their welfare at the expense of others. But inasmuch as someone who is an efficient satisfaction machine is merely someone who possesses modest preferences, satisfying her preferences first will tend to leave more resources available for satisfying others' preferences, and so point us back in the direction of a more egalitarian distribution of goods and welfare.

It should be clear that I am not suggesting that the principle of utility is really a principle of strict equality dressed up in welfarist clothing or anything like that. Obviously the question of precisely how egalitarian a strictly utilitarian distribution will be, in terms of both goods and welfare, depends on exactly what resources are available for satisfying prefererces, as well as on the

details of individuals' utility functions. (E.g., given a fairly rich but still limited stock of resources, people with very modest preferences might receive substantially less in the way of goods than others, and simultaneously enjoy substantially higher levels of welfare.) My point is rather that confusions regarding welfare and interpersonal comparisons have prompted many theorists to radically misconceive the distributive implications of utilitarianism. An enormously more efficient satisfaction machine is not someone for the utilitarian to fear. On the contrary, she is a most welcome addition to the utility calculus, and the more efficient the merrier, since her modest preferences tend to leave more resources available for satisfying the immodest preferences of others.

Similar considerations apply to "utility monster" objections to utilitarianism. Most philosophers have heard of the (possible) existence of such creatures, I take it, though printed discussion of them appears to be quite rare.[20] It may be that some theorists have more-or-less the same vague possibility in mind when talking about

---

[20] I have been unable to locate any explicit references to these exotic creatures, with the exception of the following two sentences from Nozick:

> Utilitarian theory is embarrassed by the possibility of utility monsters who get enormously greater gains in utility from any sacrifice of others than these others lose. For, unacceptably, the theory seems to require that we all be sacrificed in the monster's maw, in order to increase total utility. (Anarchy, State, and Utopia, p. 41)

utility monsters that Hart does when speaking of more efficient satisfaction machines. It is possible to put a somewhat different gloss on the notion, however, and doing so will allow us to focus on some aspects of indirect comparisons which have not yet explicitly emerged.

Intuitively, a utility monster is supposed to be someone with an insatiable appetite for various resources -- a person who will continue to derive satisfaction from indefinitely many increments of some or all goods, her "utiles" piling up without limit. Utilitarianism, the objection goes, would have us sacrifice the welfare of everyone else in society to increase the satisfaction of such a person. We can render this idea somewhat more precise as follows. Let us understand a utility monster to be someone for whom the principle of diminishing utility does not apply for a certain good G within a fairly wide range. More particularly, let us take a utility monster to be someone (i) who will receive more satisfaction from an initial increment of G than anyone else would; and (ii) for whom the principle of diminishing marginal utility does not apply with respect to at least n increments of G, where n is at least as great as the total stock of G which the community is in a position to distribute. G may be a fairly specific good (melba toast, say), or it may be something rather more general (nourishment, perhaps).

There is nothing in the present account of welfare or

interpersonal comparisons which rules out the existence of utility monsters as just described. And of course, if such a utility monster did exist, the principle of utility would indeed recommend giving every available increment of G to her, and none to anyone else.[21] This cannot reasonably be taken to be an objection to utilitarianism, however. For while indirect utility comparisons do not rule out the possibility of utility monsters in this sense, they do force us to interpret their existence in a very specific fashion.

A utility monster will receive more satisfaction from each available increment of G than anyone else would. Now notice, first of all, that the satisfaction she receives from these increments of G must make up some portion of her total possible welfare; we may assume, without loss of generality for what follows, that it comprises her entire possible welfare. The second thing to notice is that, given that the increments of G in question make up some fixed proportion of our utility monster's possible welfare, it follows that the more increments of G the community is in a position to distribute, the less satisfaction she must be understood to receive from each increment. E.g., if the community has 1000 units of G at its disposal, then (assuming that the monster does not actually exhibit increasing

---

[21]There is a "separability" assumption embedded in this conclusion which I shall not take the space to articulate.

marginal utility for G at some point in this range) each unit must count for only 1/1000 of her possible welfare.

Thus if someone really would receive as much satisfaction from the nth increment of a good as she did from the first, the present account of welfare insures that the first increment will play a correspondingly small role in her subjective valuations, with the actual extent of the correspondence fixed by n, as well as by the contribution which n increments of the good make to her total possible welfare.[22] It was part of our definition of a utility monster, however, that she receives more satisfaction from an initial increment of G than anyone else would. Thus increments of G must be understood to play a correspondingly smaller role in the subjective valuations of others, again with the extent of the correspondence fixed by the amount of G which society has at its disposal, and the role which G plays in the utility monster's valuations. For suitably large values of n, the existence of a utility monster as defined above would entail that everyone else in the community regards G's as practically worthless.

---

[22]Another way of expressing this point is that the proportionate construal of welfare imposes an upper bound on individuals' welfares -- the same upper bound, for purposes of interpersonal comparisons. It is precisely this upper bound which entails that diminishing marginal utility will hold, broadly speaking, for all goods. The upper bound insures, moreover, that to the extent that diminishing marginal utility does not hold for some range of a good, initial increments of the good must "shove over" or crowd together on an individual's scale of value, so to speak, in order to make room for additional ones.

This argument is perfectly general, applying to any goods and any amounts of those goods that one cares to think about. Realistically, the amounts of particular goods which are actually available will often work in conjunction with individuals' real-world utility functions to preclude the existence of a utility monster for those goods. E.g., given the amount of nourishment that is available in our world, a person who did not exhibit diminishing marginal utility over this range would have to place such a trivial value on each increment of nourishment that she would obviously not, given others' actual utility functions, value initial increments more than they do. Nevertheless, it remains true for goods in general that the more "insatiable" someone is with respect to those goods, the less others must be supposed to value them in order to arrive at the conclusion that the principle of utility would recommend giving most or all of the available stock to the original person.

Thus our poor utility monster has been misnamed. Her supposed insatiability results not from any real greediness on her part, but rather from the fact that she happens to value certain things which the rest of us value little if at all. Bigoted individuals have sometimes referred to people who diverged from societal norms in significant respects as "monsters" or worse; but I doubt that this is what critics of utilitarianism have had in mind when

raising the spectre of utility monsters. It is extremely important to realize, moreover, that the so-called utility monster is not merely a misdescribed innocent who happens to possess rather different values from most people. Her existence would actually constitute a positive benefit so far as the utility calculus is concerned. For, given that the goods for which she has an "insatiable appetite" represent some proportion of her total possible welfare, to the extent that she does place significant value on things which others don't, she thereby removes herself from the "competition" for scarce resources which others do value.

That utility monsters would represent a positive addition to the utility calculus should come as no surprise. We began with the supposition that the utility monster harbored an insatiable appetite for certain goods; she was, intuitively, someone who possessed highly immodest preferences regarding those goods. Upon reflection, however, we discovered that the utility monster is after all a kind of more efficient satisfaction machine with respect to certain goods, her efficiency stemming not from her possessing particularly modest preferences, but rather from the fact that her preferences are for things concerning which the rest of us are even less modest. Thus the utility monster does possess <u>comparatively</u> more modest preferences with respect to those goods; and, since the preferences at issue must occupy some proportion of her

total possible welfare, satisfying them will once again tend to leave more resources available for satisfying others' preferences. Hence the utility monster is, as with more efficient satisfaction machines generally, someone for the utilitarian to befriend rather than fear.

3. Equality and Indifference

The results of the previous section may give pause to some critics of utilitarianism, indicating as they do that certain standard objections to the theory depend on misconceiving its distributive implications. Other critics, however, will not be so easily dissuaded. For, however much the distributive implications of the principle of utility may have been misunderstood, it remains true that the principle gives no weight to distributive considerations per se, but ranks social policies exclusively in terms of their bearing on aggregate welfare. Thus utilitarianism will exhibit perfect indifference to policies which have the same consequences for aggregate welfare, regardless of how they would distribute welfare among individuals. And this, the recalcitrant critic will urge, is sufficient all by itself to show that utilitarianism is a bankrupt moral theory, whatever the picayune details of its distributive consequences.[23]

---

[23]Recalcitrant critics include Frankena and Williams, who both contend that the bare fact that utilitarianism is indifferent between policies which have the same bearing on

Some utilitarians have attempted to deflect this criticism a little by allowing a minimal weight to distributive considerations in their theories, while maintaining lexical priority for considerations of aggregate welfare. Thus Sidgwick suggested that equality might be allowed to break ties in the case of alternatives which have the same consequences for aggregate welfare.[24] Such manoeuvers are bound to strike one as a little ad hoc, however, and not merely because purely distributive considerations are unmotivated by fundamental utilitarian concerns. Sidgwick's suggestion seems ad hoc in the way that lexical priorizaticns in ethics inevitably do: it involves supposing that there are decisive reasons for preferring a more egalitarian distribution in the case of alternatives which are tied with respect to promoting aggregate welfare; and at the same time maintaining that those reasons, whatever they are, are of no consequence whatever in cases where one alternative would promote aggregate welfare even the tiniest bit more than another. Rather than opening up utilitarianism to this charge of ad hocery, I shall attempt to provide a direct two-part response to the recalcitrant critic. The first part draws on the results of the previous section, together with the fact that the critic

---

aggregate welfare constitutes a decisive objection; see Ethics, p. 33, and "A Critique of Utilitarianism", pp. 142-3.

[24] The Methods of Ethics, pp. 416-17.

almost invariably supplements his objection with what appears to be a self-defeating characterization of the supposedly objectionable feature. The second part of the response is a tu quoque which rests on attributing to the critic certain plausible views which he rejects at his peril.

(1) Does the fact that utilitarianism grants no independent weight to distributive considerations constitute a clear objection to the theory? The critic thinks so, but it is interesting to note that he typically cannot resist the temptation to fill in some of the details of the alternatives around which the objection revolves; and interesting, too, that the details always get elaborated in the same way. Williams writes:

> On the criterion of maximizing average utility, there is nothing to choose between any two states of society which involve the same number of people sharing in the same aggregate amount of utility, even if in one of them it is relatively evenly distributed, while in the other a very small number have a very great deal of it; and it is just silly to say that in fact there is nothing to choose here.[25]

I presume that Williams wishes to contrast the outcomes in this stark fashion in order to enlist as much intuitive support for his objection as possible. The difficulty with formulating the objection in this way, however, is that while it is certainly true that if an alternative which

---

[25] Ibid. Frankena develops the objection in essentially the same way, as, we may note, does Hart in the passage quoted at the beginning of the previous section.

would result in a few people enjoying high levels of welfare while many suffered was equivalent in terms of promoting aggregate welfare to one which would distribute welfare more evenly, the principle of utility would rank them indifferently; it is far from clear that the antecedent of this conditional is satisfiable in an intuitively compelling manner. And to the extent that this is so, the utilitarian may justly retort that it is rather less than circumspect to boast so confidently that there is in fact something to choose here.

What reasons are there for doubting that the antecedent of the conditional is satisfiable in any straightforward manner? We must, in the first place, carefully refrain from conceiving of welfare as something which can be distributed like ordinary material goods, a little for this person and a little (or a lot) for that one. Welfare (of the sort we are concerned with) is not a finely-grained kind of stuff which exists in limited supply, but is rather defined directly in terms of individuals' reflective and informed preferences regarding how the world will go. Hence there is in general no easy path from a situation in which a few people would enjoy high levels of welfare while many suffered to one in which the same quantity of aggregate welfare would be distributed equally. Altering circumstances so as to increase the satisfaction of some and decrease the satisfaction of others must be just that: a

matter of altering the material circumstances of society, presumably via redistributions of goods, either specific goods or more general ones like income and wealth.

The difficulty with the objection should now emerge clearly. In general it will not be possible, in anything approaching normal circumstances, to begin with a distribution in which a few people enjoy very high levels of welfare while most don't, and move via redistributions to a situation in which the same amount of welfare is more evenly distributed. For given individuals' utility functions more-or-less as they are, redistributing goods from the rich to the poor will typically result in a significant increase in aggregate welfare. There certainly is something to choose, both intuitively and from a theoretical standpoint on the utilitarian's view, between realistic alternatives on which a few people are very well off while most are poorly off, and ones on which they are equally well off: namely, the fact that the latter alternatives rank far higher in terms of promoting aggregate welfare.

That the principle of utility will not sanction obviously objectionable outcomes in the actual world is, I take it, hardly news. What I am most concerned to point out here is that a short trip through logical space is considerably less likely to achieve the desired result than most critics have imagined. The discussion of the preceding section should have convinced us that intuitive coun-

terexamples to utilitarianism are much harder to come by than is commonly supposed, and that lesson is directly applicable in the present context. Thus, suppose that we turn the present problem on its head by imagining a situation in which welfare is relatively evenly distributed, and then try to figure out how to redistribute goods so as to make a few people much better off, while simultaneously preserving aggregate welfare. The presence of an almost-utility monster would do the trick, someone who receives exactly the same satisfaction from an initial increment of G as everyone else does, at their present levels of welfare, and who exhibits constant marginal utility for as many increments of G as there are people in the community who presently possess at least one increment. We noted above, however, that the existence of utility monsters is precluded for reasonably abundant goods which people in general value to any extent, and the same applies to almost-utility monsters; the only way to render a utility monster of any stripe appreciably "insatiable" is to suppose that others don't care very much for the good or goods in question. Now, is it so clearly objectionable, beginning with a relatively egalitarian distribution of welfare, to transfer an increment of something which by hypothesis the losers don't care very much about to someone who will thereby end up very well off? If the initially egalitarian distribution was one on which everyone was very

poorly off (which it would have to be in order to arrive at an outcome in which someone was very well off while most suffered), is it inconceivable that we (not to mention those whose already miserable existence was only slightly worsened) might be a little heartened by the fact that the world was unfolding as it should in the estimation of at least one individual? Would it make a difference if the initial distribution was one on which everyone was already relatively well off; would we really look so favorably on the complaints of those who remained well off, though slightly less so, if they begrudged the lucky individual her higher level of welfare on that basis?

None of this is at all conclusive, nor is it intended to be; appeals to moral intuitions are by nature inconclusive. (For what it's worth, the string of questions above was not uttered in a purely rhetorical tone of voice; my own intuitions do not deliver very clear answers to them one way or another). My aim here is not to conclusively refute the recalcitrant critic, but rather to indicate that his objection is impotent in the absence of some specification of circumstances, in which the utilitarian's indifference to equality <u>per se</u> would make a difference, which is at least sufficiently detailed to allow one's intuitions to get a toe-hold; and to indicate further that trustworthy toe-holds of the requisite sort are not so easy to come by. A utilitarianism based on indirect interpersonal compar-

isons will favor substantial equality in actual and possible situations. (The theory is not strictly egalitarian, to be sure, at least not in most possible worlds; but then, enough egalitarianism is surely enough.) If it is true that the principle of utility sometimes refuses to choose between outcomes in which aggregate welfare is evenly distributed and ones in which "a very small number have a very great deal of it", it is also true that there is a real and insurmountable limit on how much welfare any one individual can enjoy in relation to others. While this does not entail that the situations envisaged by the critic cannot arise, we have as yet been offered no evidence that they could arise in a clearly objectionable manner. Thus if utilitarianism would in fact warrant objectionably inegalitarian outcomes in some possible circumstances, the burden remains on the critic to show that this is so.

(2) It seems reasonable to ask the critic what principle of distribution he would favor in place of the principle of utility. I shall presume in the following discussion that he is not likely to favor a principle of strict equality of distribution, of either goods or welfare, since the former seems to largely miss the point of most theorists' intuitions regarding equality, while the latter would require foregoing disturbingly many opportunities to increase individuals' welfares, simply in virtue of the fact that there are some people who are not that

well off, and who cannot be made much better off no matter what. I shall also assume that the critic is not likely to favor a version of the difference principle, Rawls' impressive argument notwithstanding, on grounds that there is no reason why people in the original position should be inclined to think, any more than most moral theorists do, that any possible benefits which might accrue to individuals whose expectations were minimally greater than those of the least advantaged members of society would be of no consequence whatever in relation to even the tiniest gains for the latter.[26]

---

[26]Note that the counterintuitive results of the principle of strict equality and the difference principle are not mere fanciful possibilities; they apply with full force in the world as it actually is. Whereas the principle of utility recommends substantial equality in the actual world, and is (I think) as egalitarian as one could hope to justify on the basis of moral intuitions in most if not all possible situations, the converse does not hold: principles which focus on promoting equality per se are enormously inefficient when it comes to promoting what individuals actually value, and one must travel a fair distance through logical space in order to erase their counterintuitive consequences. E.g., there can be little doubt that improving the real expectations of representative least advantaged persons, in whatever way possible, no matter how minimal the improvement, would be enormously expensive in any actual society -- prohibitively expensive, were it not for the fact that philosophers are always free to beggar large numbers of people in theory if not in practice. (Don't say: "Yes, but Rawls' version of the difference principle applies to expectations for primary goods, not real expectations." That is hardly a recommendation, since it suggests if anything that the POP's should stop thinking in terms of primary goods and start thinking about what really matters. In any case, this way of conveniently forgetting that some people are so lucky or unlucky in the natural lottery as to require considerably less or more goods than others in order to appreciably raise their real expectations does not differ much in the

Thus theorists who experience the pull of egalitarian intuitions, and who think that purely welfarist considerations cannot sufficiently account for the pull, are likely to favor a theory which grants independent weight to considerations of both utility and equality, while allowing lexical priority to neither of them. Paul Weirich has defended such a theory in "Utility Tempered with Equality".[27] Weirich's principle of distribution calls for first weighting individuals' true utilities in strict inverse proportion to their levels of welfare, and then maximizing aggregate weighted utility so defined. He refers to the resulting theory as "weighted utilitarianism", and argues that it treads a happy path between the counterintuitive implications of the principle of utility on one hand, and those of the difference principle[28] on the other. There are a few places in which Weirich's development of the weighted theory runs astray, but the details need not concern us. The general point is that some account similar to his, on which welfare gains for the less advantaged receive somewhat more weight than comparable gains for the more advantaged, should find favor with those who think that the principle of utility places too much

---

final analysis from explicitly abandoning the actual world in an effort to erase the counterintuitive consequences of the difference principle.)

[27]Nous Vol. XVII, No. 3 (September 1983), pp. 423-39.

[28]As well as Nicholas Rescher's "effective average" principle; see Distributive Justice (New York: Bobbs-Merrill Inc., 1966).

emphasis on welfare at the expense of equality.

Any theory of this sort is committed to holding that if two alternatives would have the same appropriately weighted consequences for utility and equality, they should be ranked indifferently. According to theories of this sort, welfare gains for the less advantaged count for somewhat more than gains to the more advantaged, but not infinitely more. Thus small gains for the less advantaged may be outweighed by larger gains for those more advantaged, and hence it will in principle be possible to construct scenarios on which pairs of alternatives have the same consequences with respect to a weighted utility/equality vector, but differ radically in how equally they distribute welfare. Of course, the mere possibility of constructing such scenarios is no more likely to show that a weighted principle would sanction objectionably inegalitarian outcomes than the corresponding possibility is for the principle of utility; somewhat less so, we may presume. Nevertheless, if the bare fact that a theory might in appropriate circumstances exhibit indifference towards alternatives which differ markedly in how equally they distribute welfare is sufficient to show that the theory is bankrupt, the egalitarian as well as the utilitarian had best start gearing up for a going-out-of-business sale.

"Yes," the egalitarian will object, "but I have a justification for remaining indifferent to alternatives

which don't differ in their utility/equality quotient. The greater inequality of one of my alternatives is offset by the greater aggregate welfare of the other. You have no such justification, since by hypothesis the outcomes of your alternatives are identical so far as aggregate welfare is concerned."

Just so. BUT -- this is critical for a general understanding utilitarianism, as well as for the present point -- this is not a justification which anyone should accept. The utilitarian does not ever appeal directly to aggregate welfare as a basic justification for anything. As critics are fond of pointing out,[29] aggregate welfare is not self-evidently valuable; nor can the aggregate of value in a community be identified with what particular members of the community value (except in the case of a community of one, in which case egalitarian concerns are hardly an issue). Sensible utilitarians keep these facts firmly in mind ("ideal utilitarians" may be a different story) and rank alternatives only by balancing the distinct gains and losses of different individuals. Thus an appeal to aggregate welfare can serve for the utilitarian only as a derived justification, its suitability in this regard stemming from the fact that one alternative will promote aggregate welfare more than another just in case it would result in gains for some which outweigh the losses to

---

[29]Cf. Hart, ibid., pp. 830-31.

others. The utilitarian certainly does have a justification, in this regard, for being indifferent to alternatives which promote aggregate welfare equally well: The lower expectations of some, on one of the alternatives, are precisely offset by the higher expectations of others.

To summarize, the egalitarian cannot appeal directly to greater aggregate welfare as a reason for displaying indifference towards a more egalitarian alternative without going doubly intuitionistic, claiming in effect that there are at least two things which have moral value independently of whether anyone values them, namely equality and aggregate welfare. Assuming that the egalitarian critic is not a blanket intuitionist of this sort, he can of course justify his indifference to the more egalitarian alternative in the same way the utilitarian would, by maintaining that the lower expectations of some individuals on the less egalitarian alternative are offset by the higher expectations of others. The difference between the two justifications will simply be that the utilitarian thinks that a loss to one individual may be offset by any comparable gain for another, whereas the egalitarian thinks that comparatively larger gains for the more advantaged are required to offset losses to those less advantaged.[30] As

---

[30]Cf.: "The utilitarian thinks that the positive effect of a gain for one individual is negated by any comparable loss to another, whereas the egalitarian thinks that the positive effect of a gain for someone less advantaged is negated only by a larger loss to someone more

noted, this difference has the effect of making it somewhat less likely that a weighted principle would sanction objectionably inegalitarian distributions of welfare in some circumstances than the unweighted principle of utility would. By the same token, however, a weighted principle will be somewhat more likely than the principle of utility to sanction objectionably egalitarian outcomes. And make no mistake, there are such outcomes; the principle of strict equality and the difference principle provide us with examples.

It should now be clear why it is so vitally important, if one is going to pay substantial heed to intuitions in moral theorizing, to get one's facts about welfare and interpersonal comparisons straight. No-one has yet demonstrated, in light of the account of individual welfare and interpersonal utility comparisons defended in the previous chapter, that acting on the advice of an unrestrained principle of utility could in fact lead to intuitively repugnant outcomes in some circumstances; all of the evidence points in the direction of thinking that critics who suppose this have systematically misinterpreted the principle's distributive implications. The fact that it is these same critics who favor restraining the principle of utility by considerations of equality should prompt the

---

advantaged." Does this equivalent description affect what our intuitions tell us about the two justifications?

suspicion that a weighted principle would run the real risk of commending intuitively repugnant outcomes on the other side. If we must risk erring on one side or the other, the safe side in my view is the one which rejects intuition as a final arbiter in moral theorizing.

4. Utility and Needs

To this point we have been discussing utilitarianism's distributive implications in a very abstract way; we have been assuming that individuals' welfares are theoretically measurable and comparable, in accordance with the account developed in the previous chapter, and exploring the distributive consequences of the principle of utility in light of that assumption. There is no question, however, that we are not and probably never will be in a position to perform the detailed measurements which would be required to directly implement the principle of utility on a comprehensive scale. Indeed, it is doubtful that we will ever be in a position to accurately measure the extent of a single individual's preferences for more than a handful of outcomes, even if we take those outcomes to be something less than the maximal ones which entered into the definitions of proportionate welfare and indirect interpersonal comparisons. If the principle of utility is to serve as a useful guide to real-world policy making, it will have to do so in some way other than by issuing detailed summaries of

expectable aggregate welfare for each available alternative.

In practice we do find ourselves estimating the relative strengths of different individuals' preferences, without relying on anything like a formal measurement procedure. In the case of people we are reasonably familiar with, we may even advance hypotheses concerning the strengths of fairly specific preferences with some degree of confidence. I'm reasonably certain, e.g., that some of my friends prefer single malt scotches to cheap blends more than others, though I've never actually tried out the experiment of giving them a cheap blend and then offering them a series of lotteries over another pair of tumblers, one filled with single malt and the other with water. In making these rough estimates of the relative importance which different people attach to different outcomes, we presumably rely heavily on observations of their behaviour, including their verbal and other expressive behaviour. Such ordinal comparisons of preference intensity are about as good as we can hope for in ordinary life; attempting much finer discriminations in either preference strengths or outcomes typically results in the confidence level of our judgments dropping off dramatically. (I may be sure that A prefers single malts to blends more than B does, but not very sure how much more; and I may not even hazard to guess whether A prefers blend X to blend Y more or less

than B does.) Still, vague as they are, such comparisons may provide useful information to the host whose supply of single malt scotch is running low.

The less familiar we are with particular individuals, and hence the fewer our opportunities have been for observing their behaviour in various circumstances, the fewer and more coarsely-grained are the judgments of interpersonal welfare that we are in a position to make with respect to them. Even in the limiting case of utter unfamiliarity, however, we do habitually compare the strengths of some preferences with a high degree of confidence. I doubt that anyone seriously suspects, e.g., that having a warm place to sleep and enough to eat doesn't mean more to people in general than having a fancier car or a color TV does, at least not in anything approaching normal circumstances. Such judgments presumably depend on certain general facts about humans and the environments (normal circumstances) in which they live, perhaps to some extent extrapolated from observing the behaviour of particular individuals with whom we're familiar. These judgments too are mostly ordinal in character. As with more discriminating judgments concerning the preference strengths of intimates, however, the mere fact that these judgments are somewhat vague does not undermine their usefulness. So long as we can be reasonably assured that having enough to eat means more to individuals generally than having a color

TV does, we can be reasonably assured that policies directed towards insuring that people have enough to eat will better serve aggregate welfare than policies directed towards putting a color TV in every living room. The fact that we are unable to say precisely how much better, with any degree of confidence, is of little or no practical significance.

No doubt this is all exceedingly obvious. One will sometimes hear it opined, however, that utilitarianism makes enormous demands on our ability to gather information concerning individuals' utility functions.[31] The suggestion appears to be that even if utilitarianism was theoretically defensible (which, it is opined, it obviously isn't), the staggering difficulties involved in applying the theory would render discussion of it pointless outside of the study or parlor. In light of such opinions, it is perhaps worth laboring the exceedingly obvious, and hence I shall take this opportunity to do so. It is perfectly obvious, apart from scepticism with respect to interpersonal utility comparisons, that some things mean more to some individuals than other things mean to others. It is furthermore sometimes obvious that some things mean more to individuals generally than other things do. (I don't mean merely that there are certain things which every individual values more than others; the claim is rather that it is a

---

[31] See e.g. Williams, ibid., p. 137.

universal or near-universal truth that, in normal circumstances, any given individual will value certain things more than any other individual will value certain other things.) Hence it is at least sometimes obvious than some policies will better promote aggregate welfare than others, and this is sufficient to provide us with a partial ranking of alternatives in accordance with the principle of utility. All of this follows simply from the ordinary judgments of interpersonal welfare which we can and do make, judgments which, given their modest character and general scepticism concerning interpersonal comparisons to the side, there is no reason to doubt the veracity of.

The principle of utility is certainly capable, then, of delivering at least some clear practical advice in real-world settings. The advice is clearest in cases where our ordinary judgments of interpersonal welfare are clearest, and these cases seem to be ones which center roughly on the satisfaction of individuals' needs. I have in mind here the general sorts of things, like adequate food, shelter, and clothing, which one points to a lack of when referring to the "needy" individuals who unfortunately persist in every society, and not the kind of thing at issue when a teenager says "But Mom, I really <u>need</u> the car tonight." Needs of the former sort are sometimes distinguished in ordinary speech from "mere preferences"; David Braybrooke refers to them as "course-of-life needs", and provisionally

adopts the notion of "functioning without derangement" in the four roles of parent, householder, worker, and citizen as their criterion.[32]

It is commonly recognized that needs are ill-suited for playing a foundational role in ethical theorizing, simply because the concept of needs itself is relational: Strictly speaking one cannot have a need for something x simpliciter, one only needs x for something else y. Thus whatever normative force needs have must apparently be derived from the things for which they are needed, and hence most of the tricky work involved in defending a theory of needs goes into (i) delimiting the y's for which genuine needs are needed in a way which will distinguish them from adventitious or spurious needs which should intuitively have no bearing on social policy (or at least, no positive bearing; e.g., the drug dealer's need for good weapons with which to hold the police at bay); and (ii) defending this criterion of genuine needs in a way which goes beyond observing that it is more-or-less extensionally adequate, so far as our intuitions concerning the importance of needs are concerned. The second part of the project is of course the trickier one. Rather than entering into a detailed exploration of the difficulties involved here, I want to suggest that the importance of

---

[32]Meeting Needs (Princeton, N.J.: Princeton University Press, 1987).

genuine needs should be understood not as deriving from the importance of the things for which they are needed, but rather as stemming directly from our ordinary judgments of interpersonal welfare, in conjunction with the principle of utility. My hypothesis, in other words, is that the concept of needs, as it actually plays a role in ethical theorizing and in the formulation and implementation of social policy, coincides to a high degree with our ordinary judgments concerning the subjective importance of certain kinds of things to people in general, and that the moral force of the concept is straightforwardly derivable from these judgments via the principle of utility.

This way of understanding needs has, I think, three distinct advantages over the more usual construal. The first is that there is something just a little odd about saying that the normative force of needs derives from whatever it is that they are needed for. To be sure, it is usually if not always possible to specify something that genuine needs are needed for; e.g., certain of the most basic ones must be met, at least in some degree, if a person is to go on living at all. The problem is not that one can't cite such facts, but rather that it sounds odd to appeal to them in justifying the importance of needs. What one wants to say, I think, is that genuine needs have a certain force of their own, a certain "self-evidence" which exempts them from standing in need of a justification in

terms of what they are needed for, unlike adventitious needs (e.g., someone's need for fresh bat's urine to finish her experiment on biological clocks; does she really need to finish the experiment?).[33] Alternatively, we might say that it seems to make at least some sense to say that genuine needs are needs which people have simpliciter. Such intuitions, to the extent that we have them, are easily explicable on the assumption that the moral force of needs stems not from the importance of what they are needed for, but rather from our ordinary judgments of interpersonal welfare via the principle of utility. (It seems likely that our ordinary judgments will to some extent be most settled with respect to things which have a high instrumental value, no matter what someone's detailed goals or values might be. Regardless, the important point from the perspective of the principle of utility is that the things in question are valued highly by people in general, never mind what they are valued for, and this serves to explain why such things -- needs -- do not stand in need of an instrumental justification.)

Secondly, needs have a curious tendency to inflate themselves over time, as the productive capacity of a society increases. This is a particularly pressing problem for any attempt to derive the importance of needs from the

---

[33]The example is borrowed from Braybrooke, ibid., p. 30.

importance of what they are needed for, since doing justice to the concept of needs as it is actually used in social and ethical decision-making will require a criterion of needs which is similarly sensitive to variations in production. Suppose, e.g., that we characterize genuine needs in terms of the requirements for an individual's functioning at a certain minimal level of adequacy in various capacities or roles, and that this criterion is more-or-less extensionally correct so far as our present intuitions regarding needs are concerned. The criterion will have to be revised if it is to continue to do justice to people's intuitions concerning what count as genuine needs against a backdrop of increasing productive capacity, and it is not clear how this can be done in a consistent way. For if we were originally prepared to argue that functioning at a certain level of adequacy in certain capacities was of sufficient moral importance to demarcate a class of genuinely important needs, but that the things required for functioning at higher levels of adequacy weren't genuine needs at all, how can we now claim that the new needs in our recently expanded set are genuine and genuinely important?

On the hypothesis that what we are inclined to class as genuine needs are just those things which we are most confident that people in general place a high value on in relation to other things, this problem evaporates. In

fact, given plausible assumptions about the character of our ordinary judgments of interpersonal welfare, the present understanding of needs positively predicts that needs will tend to keep pace with productive capacity. Our ordinary judgments are of the form 'It is a universal or near-universal truth that any given individual will value a certain thing more than any other person values a certain other thing.' The present point is most easily illustrated in cases where the things at issue admit of quantitative distinctions, for in such cases the confidence levels of our judgments may be expected to depend heavily on the quantities involved. Suppose e.g. that we are fairly confident that any given individual will in ordinary circumstances value the first few increments of a certain good G more than anyone would value a fifth, though unsure whether a third increment means more to people in general than a fifth does to some individuals. Assuming that there are enough G's available so that everybody could have a least four, the principle of utility would sanction a policy directed towards insuring that everyone had at least two increments of G, though not a policy which insured that they had more than two increments, since so far as we know aggregate welfare might be better served by giving some people two and others five. Hence on the present understanding of needs, having two increments could be counted as a genuine need, but no more.

Notice, however, that while we might not be sure that having a third increment means more to people in general than having a fifth does to some, we might still be confident that having a third increment means more to everyone than having a sixth does to anyone. If productive capacity were then to increase to the point where everyone could have at least five increments of G, the principle of utility would clearly recommend a policy on which everyone had at least three increments. Thus on the present hypothesis, having three increments could now be counted as a genuine need. (The principle of utility would also, we may note in passing, have frowned on any policy which prior to the increase in production would have allowed some people to have six G's while others had only two, even though having more than two G's wasn't a genuine need; the egalitarian implications of utilitarianism will in general outstrip policies directed towards insuring that needs are met.) While this particular example works in terms of goods which admit of degree, it will generalize readily to any case in which increasing production brings into play judgments of interpersonal welfare which, though they may have had some clear bearing on distributive policies at the old levels of production, weren't universally applicable in those circumstances. And this, I think, is likely to be the general case. Thus, not only can an account of needs which traces their normative force directly to the prin-

ciple of utility accommodate the fact that genuine needs tend to expand over time, the account actually predicts such an expansion.

Finally, theories which attempt to derive the normative force of needs from the importance of the things for which they are needed face a difficulty in saying exactly how the force of needs is supposed to impact on social policy. Intuitively, one wants to say that the goal of meeting needs should have, if not lexical priority, at least some kind of precedence in relation to promoting the satisfaction of "mere" preferences or desires.[34] The difficulty here is that when needs and their importance are understood by reference to what they are needed for, a principle which assigns precedence of some kind to meeting needs in social planning runs a serious risk of allowing needs to run out of control. We may suppose, e.g., that access to minimally decent health care will be considered a genuine need in any reasonably developed country. Now suppose that we attribute the importance of health care to the fact that minimally good health is necessary for functioning at a certain level in certain specific or general capacities. The problem is that this justification has the form of a universal generalization: what we've implicitly asserted is that anything which is required for functioning at this level in these capacities is a genuine-

---

[34]Cf. ibid., Ch. 6.

ly important need, and since minimally good health is part of what's required, access to minimal health care provisions counts as a genuine need. It is obvious, however, that some individuals require far less in the way of health care provisions than others do in order to achieve or maintain minimally good health. Hence if anything which is required for adequate functioning in certain capacities counts as a genuine need, seriously injured &c. persons will have genuine needs which far outstrip those of robust ones. Given burgeoning medical technology, there is then a real and immanent danger that a principle which assigns precedence to meeting needs may call for devoting unconscionably large amounts of resources to maintaining some individuals at, or restoring them to, minimal levels of functioning. Sometimes, however unfortunate this may be, it is better to let people die than to invest indefinite amounts of resources in saving them.[35]

Here again a conception of needs which derives their importance directly from the principle of utility avoids the difficulty. The reason is that such a conception is by nature a comparative one, in two respects: it rests the importance of needs on our ordinary judgments of interpersonal welfare, i.e. on interpersonal utility <u>comparisons</u>; and such judgments concern the <u>relative</u> importance which people in general assign to more-or-less specific things.

---

[35] Cf. ibid., Ch. 8.

Thus, while we may be confident that certain things mean more to everyone or to virtually everyone than certain other things mean to anyone, and this will in appropriate circumstances justify designating them as genuine needs, we will not in general be confident that certain things mean more to everyone than anything else means to anyone (much less to everyone). To put the point directly in terms of goods, having at least two G's (having access to specified minimum health care provisions) may count as a genuine need on the present account; but not having however many G's (however much health care) would be required to be able to do something or function adequately in some capacity.

Thus utilitarianism is capable of underwriting a concept of needs which avoids some of the serious difficulties of more standard accounts, and, if my hypothesis that this concept is in fact the one at work when people invoke needs in formulating and implementing social policy is correct, of justifying real-world policies directed towards insuring that individuals' needs are met. As noted, however, the practical consequences of the principle of utility will generally outstrip its consequences for meeting needs. The principle will also recommend redistributions, e.g., in any case where our ordinary judgments of interpersonal welfare are sufficiently clear to support the contention that some individuals would benefit more from certain goods than the people who presently possess them, even though the needs of

all the parties have been attended to. I think that our ordinary judgments are clear enough in many such cases (obviously there is room for disagreement here; keep in mind, however, that what is at issue are our pure judgments of interpersonal welfare, uncolored by intuitions regarding individuals' "entitlements" to whatever they may acquire through the exercise of their talents or by undertaking risks and so on) to support measures such as progressive taxation; just how progressive, in light of supposed incentive effects and the like, is a matter I leave for economists and psychologists and sociologists to quarrel over. There will be cases, too, where our ordinary judgments are sufficiently clear with respect to the values of identifiable groups in society to support measures undertaken specifically on their behalf (e.g., affordable childcare for single parents; wheelchair access to "essential services"). Of course, this will still not present us with anything approaching a comprehensive ranking of social alternatives; on many issues, our ordinary judgments of interpersonal welfare and/or our best guesses as to the effects of available policies are insufficiently determinate to bring the principle of utility directly to bear. Such issues are probably best left up to the market to decide, on grounds that given the standard operating assumptions this will at least result in a Pareto optimal allocation of whatever goods and resources our ordinary

judgments do not clearly tell us belong in a definite place.

My conclusion is that utilitarianism has a good deal of practical advice to offer to real-world policy makers. It will advise them in the first place to attend to individuals' needs, since policies which would result in some persons' needs going unmet are ones which are failing to capitalize on clear opportunities to increase aggregate welfare. Beyond this, it will tell them that some relatively clear opportunities exist to increase the welfare of identifiable groups without incurring an equivalent decrease in the welfare of others, and that in the absence of very strong incentive effects, progressive taxation is the appropriate way to achieve this. It will also tell them, negatively, that there are areas in which they simply cannot justify formulating and implementing policy, since in these cases there is no way of telling which policy will best promote aggregate welfare, and in such cases the market can at least be trusted (at any rate, far more than the policy makers themselves can be trusted) to hit on a Pareto optimal outcome.[36] This is a fair bit of cogent

---

[36]The importance of this negative advice should not be underestimated. Governments in developed countries squander astonishingly large amounts of resources on things like propping up failing plants and industries, when there is not the least reason to believe that doing so will promote much more than vote-buying. I doubt very much that progressive taxation and similar measures would meet with anything like the resistance they often do, were it not so obvious to those who thereby end up less well off that no

advice, I think; as much as any general ethical theory can be expected to offer, and as much as any does.

---

attempt is made or could be made to justify their loss by reference to an increase in aggregate welfare.

CHAPTER 5: UTILITY AND OBLIGATION

(The Good and the right)

Utilitarianism's central concern, I have maintained throughout, is with relating the moral worth of actions, intentions, dispositions, policies, institutions, etc. to their consequences for aggregate welfare. Consequentialist theories in general, I think, should be understood in this way; viz. as attempts to morally rank various items, notably actions, by reference to their consequences. Utilitarianism is then the particular consequentialist theory which ranks actions &c. specifically in accordance with their tendency to promote aggregate utility.

It must be admitted that this characterization of consequentialist theories in general and utilitarianism in particular controverts received opinion on these matters. An overwhelming preponderance of contemporary theorists, both those who call themselves consequentialists and those who don't, take the view that a consequentialist or teleological theory is by definition a theory which holds that an action is right if and only if it maximizes some antecedently specified good; and consequently, that utilitarianism is by definition the theory which holds that an action is right if and only if it maximizes aggregate expected utility. On the received view, in other words, consequentialist theories are by definition <u>theories of</u>

obligation.

In the final analysis, nothing much hangs on definitions themselves. If consequentialist theories are definitionally ones which equate the right with maximizing the good, then the theory which I have been defending will not be a consequentialist one, and none the worse for it; and if utilitarianism is in turn a consequentialist theory, then neither is the theory I have been defending utilitarian. Definitions aim to provide us with conceptual economy in theorizing, however, and in doing so they have the power to mislead if they are framed in ways which cause us to overlook important distinctions and possibilities. The received definition of consequentialist theories is in my view as poorly framed as they come. In recommending the conceptual shortcut of treating consequentialist theories as theories of obligation, it foists on the utilitarian an immediate concern for deontological matters which is fundamentally at odds with the original motivations for the theory, and which cannot be directly accommodated to the latter without doing serious violence to the theory as a whole; and in opposing itself to deontological theories, the definition obscures the possibility of a truly non-deontological theory, one which refuses to countenance deontological concepts of any sort on a par with axiological ones. The violence done to utilitarianism by foisting upon it an excessive concern with deontological matters

shows up most clearly in the litany of complaints that those who take the recommended shortcut have brought against the theory in recent years: that it is too demanding; that it allows no room for supererogatory actions; that it fails to take persons seriously as persons, riding rough-shod over their "personal integrity". The deterrent, I shall argue, is to cleanly separate the utilitarian's central account of the moral value of actions from her account of obligation, and relegate the latter to a distinctly subsidiary role.

It is interesting to note that the received definition of utilitarianism is quite at odds with Bentham's presentation in the Principles. The opening pages of that work make it abundantly clear that he conceived of utilitarianism in the first instance as a theory by which to assess the moral worth of actions, and not as a theory of obligation. Deontological concepts are explicitly introduced ten paragraphs on, in a passage labelled "Ought, ought not, right and wrong, &c. how to be understood", well after the principle of utility has been introduced and a number of subsidiary notions defined in terms of it:

> Of an action that is conformable to the principle of utility one may always say either that it is one that ought to be done, or at least that it is nct one that ought not to be done. One may say also, that it is right it should be done; at least that it is not wrong it should be done: that it is a right action; at least that it is not a wrong action. When thus interpreted, the words ought, and right and wrong, and others of that stamp, have a meaning; when otherwise, they

have none.[1]

The qualifications in this passage, if nothing else, leave little room to doubt that Bentham did not suppose there to be any very simple or direct connection between utility and obligation. His point here is simply that utility is the "standard of right and wrong" in the sense that deontological concepts must somehow be linked to the consequences of actions for aggregate welfare in order to be meaningful at all.

The current practice of conceiving of utilitarianism as a theory of obligation finds its seeds in Mill, whose presentation of the principle of utility contrasts sharply with Bentham's:[2]

> The creed which accepts as the foundation of morals, Utility, or the Greatest Happiness Principle, holds that actions are right in proportion as they tend to promote happiness, wrong as they tend to produce the reverse of happiness.[3]

This is a curious and highly instructive passage, inasmuch as it reveals something about how the transition from conceiving of utilitarianism as a theory of moral value to a theory of obligation may have occurred. Deontological concepts such as 'right' do not possess comparatives or superlatives, and hence do not properly admit of degree.

---

[1] An Introduction to the Principles of Morals and Legislation, p. 4.

[2] See Ch. 2, p. 12 above.

[3] Utilitarianism, On Liberty, and Considerations on Representative Government, p. 6.

There is no 'righter' or 'wronger', 'oughtest' and 'ought not-est'; actions are either right or wrong or neither, one ought either to do them or refrain from doing them or neither, and that (apart from potential qualifications or excuses, or instances where some "prima facie" obligations are overridden by others) is the end of the matter.[4] Mill, however, explicitly relies on a graded notion of right and wrong in his definition of utilitarianism: "actions are right in proportion as they tend to promote happiness".

It is difficult to be sure what Mill's intentions were in employing deviant versions of deontological concepts in his definition of utilitarianism. I think that there are clear indications that he would not have accepted the contemporary gloss of utilitarianism as a theory of obligation, but I shall not attempt to argue the point here. What is clearer, I think, is that his deviant usage was bound to, and did in fact, invite a good deal of confusion. E.g., there is no remotely plausible way to interpret the famous inference in Mill's "proof", from the fact that each person's happiness is a good to that person, to the claim that the general happiness is a good to the aggregate of all persons, in deontological terms; when Mill claims that the general happiness is a good to the aggregate of per-

---

[4]It is perhaps a testimony to the unwieldiness of such concepts that philosophers have often been led to speak of the "rightness" of actions; as though lengthening the word in this way would somehow legitimize speaking of degrees, by analogy with properties which do admit of degree.

sons, we must interpret him to mean exactly that.[5] Yet Sidgwick explicitly took Mill to be attempting to show not merely that the general happiness is a (the) general good, but that each individual ought to desire the general happiness,[6] and not without reason: Chapter IV of Utilitarianism purports to be providing whatever proof the principle of utility is "susceptible of", and Mill had earlier defined the principle with reference to 'right' and 'wrong'. (Unfortunately, Sidgwick seems to have overlooked the fact that Mill was not using these terms in their usual sense. Given the deviant usage, Sidgwick's interpretation would have Mill attempting to show, at best, that it was "righter" (more right?; displayed more rightness?) to desire the general happiness. More likely, I think, Mill had consciously or unconsciously dropped any reference to deontological concepts for the purposes of the proof, and offered it for what it is, namely an argument against intuitionistic conceptions of moral value.)

Whatever Mill's intentions were in employing deviant versions of deontological concepts, by the time Sidgwick got around to writing The Methods of Ethics the damage had been done. For Sidgwick, utilitarianism is essentially a

---

[5]Ibid., p. 33. This part of the proof bears an obvious resemblance to Bentham's argument that, since a community is simply a collection of individuals, the interest of the community is just the "sum of the interests of the several members who compose it"; see Bentham, ibid., p. 3.

[6]The Methods of Ethics, p. 388.

theory of obligation, a theory which states that each individual ought to desire the general happiness. The entire architectonic of the Methods is shaped around this presumption, right down to the final conclusion that the doctrine must ultimately rest on an appeal to intuition. And certainly the deontological concepts employed therein are in no way deviant; utilitarianism imposes on individuals a duty to maximize aggregate welfare, period.

The mistake in all this, I want to suggest, is to suppose that a moral theory should be centrally preoccupied with right and wrong in the first place. Deontological concepts are most certainly moral concepts, and a theory which didn't have anything to say about their application would be no moral theory at all. It does not follow, however, that we should treat them as foundational moral concepts, and this is precisely the mistake which Mill encouraged, and which Sidgwick finalized. In my view, deontological concepts can and must take a back seat to the utilitarian's account of moral value. Indeed, I suspect that in order to make real sense of the complexities of moral decision-making, deontological concepts must take a back seat in any ethical theory to an account of the moral value of actions. The right is not, as is commonly supposed, one of a pair of fundamental ethical concepts.

There is a prima facie difficulty involved in any attempt to link up deontological concepts with axiological

ones, given that the latter are graded concepts and the former are not. The simplest way to connect the two would appear to be to fix the non-graded concepts at the limit of the graded ones; i.e., to equate 'right' with 'best'. This yields, of course, the usual definition of consequentialism, along with the usual flurry of objections. Is there a better way for the consequentialist in general and the utilitarian in particular to connect deontological notions to basic values? One natural suggestion, given that equating the right with the best is apparently too stringent, is to adopt something like a "satisficing" account, on which right actions don't necessarily have to be the best available actions, they just have to be "good enough". Michael Slote has recently suggested an account of this sort in Beyond Optimizing.[7] While his account ultimately founders in important respects, examining it briefly will help us see our way clear to finally arriving at an adequate utilitarian account of obligation.

Slote's argument begins with the observation that there is a strong parallel between contemporary theories of individual rationality and consequentialist moral theories, as standardly defined. According to a majority of decision

---

[7]Beyond Optimizing: A Study of Rational Choice (Cambridge, Mass.: Harvard University Press, 1989). See also his "Satisficing Consequentialism", Proceedings of the Aristotelian Society, Supplementary Volume LVIII (1984), pp. 139-63, as well as Philip Pettit's response to the latter, ibid., pp. 165-76.

theorists, rational individuals are ones who maximize their expectation of personal good. Similarly, consequentialist moral theories are ones which require individuals to maximize the expectation of some impersonal good, typically aggregate welfare somehow conceived. The points at which consequentialist moral theories conflict with our common-sense view of morality have been thoroughly mapped in recent years,[8] and all roads lead to the conclusion that (i) they place too many demands on individuals, and (ii, more contentiously) that there are intuitive constraints on what one may do in the name of maximizing an impersonal good which are not reflected in standard consequentialist theories (e.g., killing someone to save others).

What has been less frequently noticed or totally ignored, according to Slote, is that we also possess a common-sense view of rational decision-making, a kind of "folk theory" of individual rationality, which is at odds with maximizing theories of rationality in just the way that folk morality stands opposed to standard consequentialist theories. On Slote's view, our folk theory of individual rationality does not require rational persons to continually maximize their expected good, and may even hold that it is sometimes irrational to so maximize; folk rationality is a kind of satisficing theory which permits

---

[8]See e.g. Slote, Common-Sense Morality and Consequentialism (Boston: Routledge and Kegan Paul, 1985).

people to aim for the "good enough", and may actually require them to moderate their pursuit of the good when they have achieved something good enough, though still short of the best. Given that common sense finds fault with both maximizing accounts of individual rationality and maximizing consequentialist moralities, in more-or-less the same way, the mutual support which our folk theories of rationality and morality provide for each other may in turn provide some support for the view that the standard theories should be supplanted by satisficing versions: satisficing rationality on the one hand, and satisficing consequentialism on the other.

It is important to realize that the concept of satisficing which Slote employs differs markedly from the one originally developed by Herbert Simon.[9] The latter concept is intended to explain rational choice in situations where there are costs involved in decision-making, and where there is uncertainty as to whether an evaluation of all of the available alternatives would repay the costs involved, or perhaps even uncertainty as to what the available alternatives are. In such situations many decision-makers apparently form some conception of what a "good enough" outcome would be, and then search through or search out

---

[9]See e.g. "A Behavioural Model of Rational Choice", Quarterly Journal of Economics 69 (1955), pp. 99-118, and "Theories of Decision Making in Economics and Behavioural Science", American Economic Review XLIX (1959), pp. 253-83.

alternatives until they arrive at one with a result which is at least good enough (if the result happens to be better than good enough, so much the better). In other words, they satisfice rather than maximize, in situations where (because of uncertainty) standard expected-utility maximization is undefined. Such behaviour is common, and seems intuitively rational, or at least not irrational. The important point is that on Simon's conception, agents have an apparent reason to satisfice, given the costs of decision-making, though the uncertainty involved does not permit us to formulate this reason in standard utility-maximizing terms.

Slote, however, contends that on our folk theory of individual rationality, it is sometimes rational for an agent to settle for a good enough outcome even when there is a better outcome achievable, all things considered, and the agent knows it. His support for this claim comes mainly from a series of examples which he hopes will provoke favourable intuitions. The examples all concern instances of someone turning down a supposedly certain or near-certain good thing, in favor of something less good but apparently good enough. One concerns, e.g., a person who has had a good lunch and is not now hungry, who knows that there are candy bars and drinks in the fridge next to her desk, stocked gratis by the company, who knows that she would enjoy such a snack, and who is not worried about her

diet or spoiling her dinner or the minimal effort involved in going to the fridge or anything like that. Slote contends that such a person might nevertheless refrain from getting a snack, and rationally do so, because she feels no need of it, being perfectly satisfied as she is.

The problem here is that it seems natural to describe the person not as rejecting a sure good thing, but as simply being indifferent to the snack, and not in any technical or refined sense of "indifference". (She is, by hypothesis, "perfectly satisfied" as she is.) Notice how odd it would sound, e.g., for the person in question to protest to the ascription of indifference with: "Oh no, you misunderstand, I do want the snack. I've thought carefully about the matter, and I'm certain that I would enjoy it. So you see, I really do prefer having the snack, in itself as well as all things considered, I'm just not going to have it." On the one hand she has confessed a reason (a personal motive) for doing something, simultaneously proclaimed that she can not think of any reason for not doing it, and yet she does not do it. Such a person, I think we are inclined to say, is either not telling the whole truth or is irrational.

Slote's other examples suffer from the same difficulty, and eventually he resorts to speaking of the satisficing persons in his examples as "preferring" to forego a supposedly sure good, as rejecting an "unwanted" though

again purportedly better outcome, as possessing second-order preferences for being moderate in their desires, and so on. He does not take this to be a refutation of his basic thesis concerning folk rationality, however, because he thinks that what individuals want or prefer, all things considered, cannot in general be identified with their personal good. The argument here is the usual one from the possibility of altruism:

> If altruism makes sense, then presumably so too does the notion of self-sacrifice. But the idea of deliberate self-sacrifice involves the assumption that what a person (most) wants need not be what advances his own personal well-being, what is (in one everyday sense) best for him. And this conceptual point carries over to discussions of moderation and satisficing. Just because the moderate individual asks for less money than he possibly could doesn't, for example, mean that additional wealth wouldn't be a good thing for him....There is conceptual space for and human understandability in the idea of a personal good or element of one's own well-being that one simply doesn't care about or wish to have -- and that one actually rejects -- because one feels well enough off without it.[10]

Something has gone decidedly wrong here, however. Perhaps there is an everyday sense in which what is best for a person need not coincide with what she most wants. And perhaps in this everyday sense of personal well-being, it is intuitively rational for people to sometimes deliberately reject what would advance their well-being. If so, that is apparently because it is intuitively or common-sensically rational for them to do what they most want to do, which

---

[10] Beyond Optimizing, pp. 18-19.

may involve rejecting what is in their everyday (i.e. common-sensical) interest. If anything, then, Slote's examples seem to show that there is a tension between what we common-sensically take to be in people's interests and what we common-sensically think it is rational for them to do, since we apparently think that it is rational for them to do what they care about, and not what is in their interest. The examples were supposed to convince us, however, that folk rationality is at odds with standard maximizing accounts of individual rationality, not with our common-sense view of what is in someone's interest; whereas what they seem to show is that the two are so far in lock-step. For standard maximizing accounts do not say that rational individuals maximize something that they don't care about or wish to have, but precisely that they maximize what they <u>do</u> care about and <u>do</u> wish to have.

The only conclusion to be drawn, I think, is that theorists have generally overlooked the common-sense satisficing view of individual rationality because there's no such thing (except in the sense that it is common-sensically rational to forego things which one doesn't desire, and this may appear to be a kind of satisficing with respect to some common-sense conception of well-being; but this is true from the perspective of explicitly maximizing accounts of rationality as well). My diagnosis of Slote's error is this: Even though the first seven

chapters of his book are devoted to discussing his hypothetical folk theory of rationality and only the last discusses its implications for ethical theorizing, it is clear that much or most of his thinking about the former has been shaped by the latter; analogies and examples drawn from the standard stock of objections to consequentialist moral theories turn up on almost every page, as a means of illuminating and advancing the folk theory. Slote has thought long and hard about the problems with standard consequentialist theories, particularly the fact that they appear to be far too stringent. Conceiving of consequentialist theories in the standard way, however, as theories of obligation, involves thinking of deontological concepts as being on a par with axiological ones, imbedded side by side within the foundations of the theory. And, having thought long and hard about these matters, Slote has rightly recognized that there are no factors internal to a consequentialist theory which would license equating the right with anything less than the best. (E.g., moving to an indirect or "rule-consequentialism" won't do the trick, because any consequentialist justification for adopting a rule is also a justification for violating it when doing so would have recognizably better consequences -- the "rule worship" problem.) He has thus been prompted to look outside of consequentialist theories for something suitably "principled" with which to restrain their deontological

excesses (suitably principled, because we are after all looking for a way to restrain fundamental moral concepts such as 'right' and 'wrong'). This something he discovers in supposed facts about the moderate or satisficing character of our common-sense views concerning rationality, facts which so far as I can see simply aren't there.

The mistake, then, is to take the shortcut of conceiving of consequentialist theories as theories of obligation, and thereby bless deontological concepts with a moral standing which they don't deserve. Refusing to take the shortcut will remove any temptation to think of deontological principles as first moral principles,[11] and thus leave us free to deploy a more adequate satisficing notion in our account of obligation. In particular, refusing to grant deontological principles the status of first principles allows us to replace Slote's unmotivated satisficing with a clearly motivated variety, since we are now free to draw on external motivational factors, rather than looking for reasons or motivations which stem directly from the principle of utility. The motivations which are especially relevant here are individuals' personal values, insofar as they diverge from the moral rankings which do issue from the principle of utility. Roughly speaking, what I want to say is that an individual has acted rightly just in case

---

[11]Which, as Sigdwick rightly maintained, would require a commitment to some form of intuitionism.

she has performed an action which is morally good enough, and wrongly just in case she has performed an action which is not morally good enough, where what counts as good enough in the circumstances will depend in part on how much better she might have done, but also, crucially, on what her own values are in relation to what she did and what she might have done.

Obligations are things which people impose on each other and on themselves via practices of praising and blaming, punishment and the like; or else they are nothing. I shall have something to say in a moment about this baldly positivistic claim, but for now it is perhaps best to treat the present thesis concerning the satisficing character of obligation simply as a hypothesis about our actual judgments concerning right and wrong, at least those made outside of the study. The hypothesis, then, is this: When we judge that someone has acted rightly, what we judge is that by our lights she has acted "well enough", where this depends not only on what better actions were available to her, but also on how her own personal goals and values relate to those actions. More specifically, in cases where an individual has less of a personal stake in the proceedings, acting rightly/well enough will, other things equal, require performing a morally better action than cases where she has a greater personal stake. For example, while we might be prepared to allow that an individual ought to

perform an action which is reasonably conducive to aggregate welfare if the cost to her is some small inconvenience, few people (perhaps only those captive to a "consequentialist" moral theory) would judge that she is obliged to perform an action equally conducive to aggregate welfare if that would require, say, sacrificing the life of a child; that will seem to most people, I expect, an unreasonable demand to make of anyone, unless the moral stakes are enormously high.[12]

The hypothesis needs to be refined somewhat. Deontological judgments as I have just characterized them are highly context-sensitive, the relevant aspects of the context being (i) the moral stakes involved, and (ii) the personal stakes of the agent in question. Our deontological judgments may cover a wide range of cases, however, from judgments concerning an individual's behaviour in a particular situation, to judgments concerning how people in general should act in situations of a specific or general kind. The more general a deontological judgment is, the less context-sensitive it must be, since there is less specific information to go on concerning the two relevant factors. Thus, in order to avoid "falsely" imputing

---

[12]Cf. Bentham, ibid., p. 20n., where in hypothetical conversation with a moral intuitionist he asserts that the latter has a moral duty to prevent people from performing actions which are detrimental to aggregate welfare "if it is what lies in your power, and can be done without too great a sacrifice".

obligations to individuals whose personal stakes in a given situation might be higher than those of people who typically find themselves in situations of that sort, general deontological judgments will generally be somewhat less stringent than more specific ones sometimes are. Legal obligations form an interesting special case. Laws must be relatively clear and simple and universally applicable in order to be effectively promulgated and enforced, and the practical impossibility of formulating a law for every situation also dictates that most laws must be quite general in scope. For these reasons and others, laws are quite insensitive to deontologically relevant aspects of the situations in which they apply (e.g. it is not a legal excuse for running a red light that there was no need to wait, since it was 3 a.m. and there was no other traffic on the road; nor is being late for an important appointment an excuse; on the other hand, self-defense is a excuse for killing someone). This explains why legal obligation is a minimal kind of obligation, specifying minimally adequate standards of behaviour for the members of a society.

Conversely, individuals' social or moral obligations will typically outstrip their legal obligations, because the context-insensitivity which is a necessary feature of law is no longer a factor, or less of a factor. When we judge on a particular occasion that someone has acted rightly or wrongly in the circumstances, we are free to

make use of whatever knowledge we may have of her personal stake in the matter, as well as knowledge of the moral stakes involved, and hence we may be prepared to impute obligations to her on this occasion that we would be disinclined to impute to people in general on occasions of this sort. Notice that if the present hypothesis concerning the satisficing character of our deontological judgments is correct, any circumstance in which an individual has nothing at stake is one in which she acts rightly if and only if she performs the best available action. This follows from the fact that the concept of satisficing at work is a motivated one, and that the motivations in question are individuals' personal values: in cases where an individual has nothing at stake, the motivation for satisficing disappears and hence the present account of obligation collapses into a maximizing one.[13] This, I

---

[13]Strictly speaking this should be modified somewhat in order to take into account Simon's original insights concerning satisficing in contexts where there is uncertainty as to whether an evaluation of all of the alternatives available to an agent would repay the costs involved. In the case of moral decision making there may be two kinds of costs involved, namely personal costs to the agent, and moral costs. The latter will typically engender a satisficing element in deontological judgments regardless of what an agent's personal stakes may be in a given instance. I.e., since there may well be non-neglible moral costs involved in determining the precise extent to which various alternatives available to an agent would promote aggregate welfare, it seems reasonable to adopt a satisficing approach to obligation in general, quite independently of whether the personal interests of agents further constrain our deontological judgments. The present point may then be expressed as the observation that in instances where an agent has no personal stake at all in a matter,

think, gets things intuitively just about right. An account which relied on unmotivated satisficing of the sort defended by Slote, on the other hand, would apparently commit one to the view that it is sometimes permissible for an individual to fail to perform a morally better action, even when doing so would not cost her so much as the effort of lifting her little finger.

We should distinguish between social or moral obligation and personal obligation. Our judgments of right and wrong are of course not restricted to the actions of others; we also make deontological judgments concerning our own behaviour -- we impose, so to speak, obligations on ourselves. Typically, I think, most people are inclined to judge their own case somewhat more stringently than that of others from a deontological point of view; i.e., we often take ourselves to have obligations which outstrip those we are prepared to impute to others in similar circumstances, and which others are prepared to impute to us. It may be that this stems from an underlying commitment to some form of liberalism, a general suspicion that it is somehow improprietous to tell others what they should and shouldn't be doing; no-one likes a moral busybody. On the other hand, our satisficing account of deontological judgments provides an explanation for our reluctance to be moral

---

she acts rightly just in case she meets or exceeds a satisficing threshold which is motivated by purely moral considerations.

busybodies. Judgments of right and wrong are sensitive both to variations in the moral worth of the actions available to agents, and to variations in their personal values in relation to those actions. Thus in making deontological judgments we must rely to some extent on estimations of different individuals' utilities; i.e., such judgments are dependent on interpersonal utility comparisons, not only for determining the moral worth of actions, but also the extent to which deontological satisficing is a factor in the circumstances. Assuming a modest form of privileged access to our own utility functions, it is easy to see why our deontological judgments should be more restrained with respect to others' behaviour than with respect to our own: we run less of a risk, in our own case, of falsely imputing obligations.

Most contemporary critics of utilitarianism appear to be of the opinion that the theory fails most dismally when it comes to issues of personal moral decision-making. Yet it is just here, I think, that an adequately formulated utilitarianism really comes into its own as an ethical theory <u>per se</u>, rather than as a theory of social choice. Consider Williams' famous story about Jim the botanist and Pedro the captain:

> Jim finds himself in the central square of a small South American town. Tied up against the wall are a row of twenty Indians, most terrified, a few defiant, in front of them several armed men in uniform. A heavy man in a sweat-stained khaki shirt turns out to be the captain in charge

> and...[he] explains that the Indians are a random group of the inhabitants who, after recent acts of protest against the government, are just about to be killed to remind other possible protesters of the advantages of not protesting. However, ...the captain is happy to offer [Jim] a guest's privilege of killing one of the Indians himself. If Jim accepts, then as a special mark of the occasion, the other Indians will be let off. Of course, if Jim refuses, then there is no special occasion, and Pedro here will do what he was about to do when Jim arrived, and kill them all.[14]

Williams went on to criticize utilitarianism for maintaining not only that Jim should kill the Indian, but for asserting so quickly and confidently that this is obviously the right answer. The problem, according to Williams, is not just that utilitarianism imposes extreme moral demands on individuals, but that it does so with total disregard for how their own values and interests bear on moral decision-making, except insofar as these play a small part in the impersonal goal of maximizing aggregate expected utility. In ignoring how individuals' personal goals and projects bear on moral decision-making, utilitarianism fails to treat persons seriously as persons, since it is incapable of making any sense of the idea of "personal integrity" in moral choice.

All of this does seem to follow if we take utilitarianism to be a theory of obligation, one which quickly and confidently asserts that individuals are obliged to maximize aggregate welfare. I have been urging, however, that

---

[14] "A Critique of Utilitarianism", p. 98.

utilitarianism is not centrally a theory of obligation, but is rather a theory of the moral value of actions; obligations come into the picture only much later, via our judgments concerning what individuals should do not only in light of the bearing of their actions on aggregate welfare, but also in light of how their personal values relate to the moral worth of their actions. Consider Jim's predicament. For my own part, I would not judge Jim to be under an obligation to kill the Indian; if he did so, that could only be deemed supererogatory. Neither would I take myself to be under an obligation, if I were in Jim's place. I do not know what I would do in such a situation. I hope that I would have the fortitude to shoot one of the Indians, but any prediction on this score would be rash, given my revulsion to the idea of deliberately killing someone who in no way represents a threat to me or to those I care about.

Crucially, however, if I did fail to shoot the Indian, I would know that I might have done better than I did, morally speaking, and this seems to me to be a critical aspect of the decision problem. It is critical precisely because it is my own action, something morally better which I might have done, which is at issue here, and not the cold and impersonal goal of maximizing aggregate utility. To put a point on it, obligations seem more-or-less irrelevant to assessing Jim's case. What is relevant is that it is

impossible to make sense of Jim's predicament, or our own predicament, if we imagine ourselves in Jim's shoes, unless we credit Jim or ourselves with a clear recognition that one of the actions available in the circumstances is of substantially greater moral worth than the other, though substantially less attractive from a personal point of view. Without this recognition there is no predicament at all, moral or otherwise. A man who advances his personal projects with no regard for the moral worth of his actions is not someone with personal integrity, he is a psychopath.

Return now to my earlier suggestion that obligations are things which we impose on each other through practices of praising and blaming, and systems of reward and punishment generally, or else they are nothing. Surely there is a difference between what the law says we should do, or what others say we should do, or what we ourselves think we should do, and what it is really right for us to do? The question is confused. There is a fact of the matter about what the law requires of us, and what others require of us, and what we require of ourselves. But there is no risk of concluding here that thinking that something is right or wrong makes it so, because there is nothing here to be right or wrong (i.e., correct or incorrect) about; there is no fact of the matter at all about what it is morally right for us to do. It does not follow, however, that any system of legal or moral or personal obligation is as good as any

other. The principle of utility is as capable of deciding the moral worth of various practices governing rewards and sanctions as it is of assessing the worth of actions generally, and hence deontological practices are morally criticizable like other practices. The confusion comes in supposing that the fact that some deontological practices are morally better than others entails that there is a fact of the matter about what our obligations are. That is simply to iterate the mistake that I have been inveighing against for most of the present chapter, to suppose that we are obliged to choose just those deontological practices which are the morally best ones in the circumstances. The principle of utility tells us what it is morally better and worse for us to do, including that there are morally better and worse ways of imposing obligations on ourselves and on others, but it does not and cannot directly impose obligations on us; to suppose that it could is a bit of superstitious nonsense ("nonsense on stilts").[15]

In sum: Utilitarianism does not impose on us a duty to be moral heroes or anything else. If it's sainthood that we aspire to, that is something which we shall have to impose on ourselves; and if we aspire to something less

---

[15]Cf. Mill's discussion of the contrast between the "transcendental moralists", who see moral obligation as "a transcendental fact, an objective reality belonging to the province of 'Things in themselves'," and the utilitarian view on which obligation is "entirely subjective, having its seat in human consciousness only." (Utilitarianism, p. 27)

than sainthood, that too i something for which we must finally accept responsibility -- we should not expect an ethical theory to directly sanction our refusal to be something morally less than we might be. Too many ethical theorists have for too long been obsessed with questions of right and wrong in the abstract, with what people are "really" obliged to do and refrain from doing quite independently of how they actually impose obligations on themselves and others. Such an obsession may have been appropriate to a time when morality was conceived of primarily in terms of the edicts of a divine being, at first somewhat harsh, and then later on a little more understanding; for then it made some sense to suppose that the divine being's judgments concerning what counted as good enough in the circumstances should override those of mere mortals. But that time has long since passed away. Utilitarianism is an ethical theory suited to the human condition, and humans impose (and refuse to impose) obligations on themselves.

We may wonder what more one could ask of an ethical theory, beyond clear and unequivocal deliverances regarding what is morally better and worse, short of some kind of moral hand-holding. That utilitarianism refuses to provide such hand-holding is a better indication than any that it, and perhaps it alone, is prepared to respect persons for what they are.

REFERENCES

Armstrong, D.M. Universals and Scientific Realism Vol. II: A Theory of Universals. Cambridge University Press, 1978.

Arrow, Kenneth J. Social Choice and Individual Values. 2nd ed. New Haven: Yale University Press, 1963.

----------. "Some Ordinalist-Utilitarian Notes on Rawls's Theory of Justice." Journal of Philosophy 70 (1973): 245-63.

----------. "Formal Theories of Social Welfare." In Dictionary of the History of Ideas, Vol. 4, pp. 276-84. Edited by Philip P. Wiener. New York: Charles Scribner's Sons, 1973.

----------. "Extended Sympathy and the Possibility of Social Choice." American Economic Review Papers and Proceedings 67 (1977): 219-25.

----------. Collected Papers Volume 1: Social Choice and Justice. Oxford: Basil Blackwell, 1984.

Arrow, Kenneth J. and Scitovsky, Tibor, eds. Readings in Welfare Economics. London: George Allen and Unwin Ltd., 1969.

Bentham, Jeremy. An Introduction to the Principles of Morals and Legislation. Oxford: Clarendon Press, 1823.

Bergson, Abram. "A Reformulation of Certain Aspects of Welfare Economics." Quarterly Journal of Economics 52 (1938): 310-34.

Brandt, Richard B. A Theory of the Good and the Right. Oxford: Clarendon Press, 1979.

----------. "Two Concepts of Utility." In The Limits of Utilitarianism, pp. 169-85. Edited by Harlan B. Miller and William H. Williams. Minneapolis: University of Minnesota Press, 1982.

Braybrooke, David. Meeting Needs. Princeton, N.J.: Princeton University Press, 1987.

Conwell, Russell H. "Acres of Diamonds." In Cases in Business and Society, 2nd ed., pp. 33-40. Edited by Scott H. Partridge. Englewood Cliffs, N.J.: Prentice Hall Inc., 1989.

d'Aspremont, C. and Gevers, L. "Equity and the Informational Basis of Collective Choice." Review of Economic Studies 44 (1977): 199-210.

Dworkin, Ronald. Taking Rights Seriously. Cambridge, Mass.: Harvard University Press, 1977.

----------. "Rights as Trumps." In Theories of Rights, pp. 153-67. Edited by Jeremy Waldron. Oxford University Press, 1984.

Edgeworth, F.Y. Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences. London: C. Kegan Paul & Co., 1881.

----------. "The Pure Theory of Taxation." In Papers Relating to Political Economy, Vol. II, pp. 63-125. London: Macmillan and Co., 1925.

Ellsberg, D. "Classic and Current Notions of 'Measurable Utility'." The Economic Journal 64 (1954): 528-56.

Fishburn, Peter C. Utility Theory for Decision Making. New York: John Wiley & Sons, 1970.

Frankena, William K. Ethics. Englewood Cliffs, N.J.: Prentice Hall, Inc., 1963.

Friedman, Milton. "Lerner on the Economics of Control." The Journal of Political Economy 55: 406-16.

Gauthier, David. "On the Refutation of Utilitarianism." In The Limits of Utilitarianism, pp. 144-63. Edited by Harlan B. Miller and William H. Williams. Minneapolis: University of Minnesota Press, 1982.

----------. Morals By Agreement. Oxford University Press, 1986.

Gibbard, Allan. "Interpersonal Comparisons: Preference, Good, and the Intrinsic Reward of Life." In Foundations of Social Choice Theory, pp. 165-93. Edited by Jon Elster and Aanund Hylland. Cambridge University Press, 1986.

Griffin, James. "Modern Utilitarianism." Revue Internationale de Philosophie 36 (1982): 331-375.

----------. Well-Being: Its Meaning, Measurement, and Moral Importance. Oxford: Clarendon Press, 1986.

----------. "Well-being and its Interpersonal Comparability." In Hare and Critics: Essays on "Moral Thinking", pp. 73-88. Edited by Douglas Seanor and N. Fotion. Oxford: Clarendon Press, 1988.

Hammond, Peter J. "Equity, Arrow's Conditions, and Rawls' Difference Principle." Econometrica 44 (1976): 793-804.

Hare, R.M. "Justice and Equality." In Justice and Economic Distribution, pp. 116-31. Edited by John Arthur and William H. Shaw. Englewood Cliffs: Prentice-Hall Inc., 1978.

----------. "Comments on Richards." In Hare and Critics: Essays on "Moral Thinking", pp. 255-60. Edited by Douglas Seanor and N. Fotion. Oxford: Clarendon Press, 1988.

Harsanyi, John C. "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility." Journal of Political Economy 63 (1955): 309-21.

----------. "Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory." American Political Science Review 59 (1975): 594-606.

----------. "Nonlinear Social Welfare Functions: Do Welfare Economists Have a Special Exemption from Bayesian Rationality?" Theory and Decision 6 (1975): 311-32.

----------. Essays on Ethics, Social Behaviour, and Scientific Explanation. Dordrecht: D. Reidel, 1976.

----------. "Rule-utilitarianism and Decision Theory." Erkenntnis 11 (1977): 25-53.

Hart, H.L.A. "Between Utility and Rights." Columbia Law Review 79 (1980): 828-46.

Hicks, J.R. and Allen, R.G.D. "A Reconsideration of the Theory of Value." Economica Vol. 1, No. 1 (February 1934): 52-76 and Vol. 1, No. 2 (May 1934): 196-219.

Jevons, W. Stanley. The Theory of Political Economy. 4th ed. London: MacMillan and Co., 1911.

Kaldor, Nicholas. "Welfare Propositions and Interpersonal Comparisons of Utility." The Economic Journal 49 (1939): 549-52.

Kim, Jaegwon. "Psychophysical Supervenience." Philosophical Studies 41 (1982): 51-70.

Krantz, David H.; Luce, R. Duncan; Suppes, Patrick; and Tversky, Amos. Foundations of Measurement. Vol. 1: Additive and Polynomial Representations. New York: Academic Press, Inc., 1971.

Lange, Oscar. "The Foundations of Welfare Economics." Econometrica 10 (1942): 215-28.

Lewis, David. "Extrinsic Properties." Philosophical Studies 44 (1983): 197-200.

----------. On the Plurality of Worlds. Oxford: Basil Blackwell Ltd., 1986.

Little, I.M.D. A Critique of Welfare Economics. Oxford University Press, 1950.

Luce, R. Duncan and Raiffa, Howard. Games and Decisions: Introduction and Critical Survey. New York: John Wiley & Sons, 1957.

MacIntosh, Duncan. "Retaliation Rationalized." Manuscript presented to the annual meetings of the Canadian Philosophical Association. Windsor, May 1988.

Mill, J.S. Utilitarianism, On Liberty, and Considerations on Representative Government. Edited by H.B. Acton. London: J.M Dent & Sons, 1972.

Miller, Harlan B. and Williams, William H., eds. The Limits of Utilitarianism. Minneapolis: University of Minnesota Press, 1982.

Nozick, Robert. Anarchy, State, and Utopia. New York: Basic Books Inc., 1974.

----------. "Interpersonal Utility Theory." Social Choice and Welfare 2 (1985): 161-79.

Overvold, Mark C. "Self-Interest and the Concept of Self-Sacrifice." Canadian Journal of Philosophy 10 (1980): 105-18.

----------. "Self Interest and Getting What You Want." In The Limits of Utilitarianism, pp. 186-94. Edited by Harlan B. Miller and William H. Williams. Minneapolis: University of Minnesota Press, 1982.

Pareto, Vilfredo. Manual of Political Economy. Translated by Ann S. Schwier. New York: Augustus M. Kelley, 1971.

Pettit, Philip. "Satisficing Consequentialism." Proceedings of the Aristotelian Society. Supplementary Volume LVIII (1984): 165-76.

Ramsey, Frank. Foundations: Essays in Philosophy, Logic, Mathematics and Economics. Atlantic Highlands, N.J.: Humanities Press Inc., 1978.

Rawls, John. A Theory of Justice. Cambridge, Mass.: Harvard University Press, 1971.

Rescher, Nicholas. Distributive Justice. New York: Bobbs-Merrill Inc., 1966.

Richards, David A.J. "Prescriptivism, Constructivism, and Rights." In Hare and Critics: Essays on "Moral Thinking", pp. 113-27. Edited by Douglas Seanor and N. Fotion. Oxford: Clarendon Press, 1988.

Robbins, Lionel. "Interpersonal Comparisons of Utility: A Comment." Economic Journal 48 (1938): 635-41.

----------. An Essay on the Nature and Significance of Economic Science. 2nd ed. London: MacMillan and Co., 1945.

----------. "Robertson on Utility and Scope." Economica 20 (February 1953): 99-111.

Rothenberg, Jerome. The Measurement of Social Welfare. Englewood Cliffs, N.J.: Prentice-Hall Inc., 1961.

Samuelson, Paul. "A Note on the Pure Theory of Consumer Behaviour." Economica 5 (1938): 61-71.

Schick, Frederic. "Beyond Utilitarianism." Journal of Philosophy Vol. LXVII, No. 20 (1971): 657-66.

Schwartz, Thomas. "Human Welfare: What It Is Not." In The Limits of Utilitarianism, pp. 195-206. Edited by Harlan B. Miller and William H. Williams. Minneapolis: University of Minnesota Press, 1982.

Scitovsky, Tibor. "A Note on Welfare Propositions in Economics." The Review of Economic Studies 9 (1941): 77-88.

Sen, Amartya K. Collective Choice and Social Welfare. San Francisco: Holden-Day, Inc., 1970.

----------. "Behaviour and the Concept of Preference." Economica 40 (1973): 241-59.

----------. "Welfare Inequalities and Rawlsian Axiomatics." In Foundational Problems in the Special Sciences, pp. 271-92. Edited by Robert E. Butts and Jaakko Hintikka. Dordrecht: D. Reidel, 1977.

----------. "Utilitarianism and Welfarism." Journal of Philosophy Vol. LXXVI, No. 9 (1979): 463-89.

----------. "Interpersonal Comparisons of Welfare." In Economics and Human Welfare: Essays in Honor of Tibor Scitovsky. Edited by M. Boskin. New York: Academic Press, 1979.

----------. Choice, Welfare, and Measurement. Cambridge, Mass.: MIT Press, 1982.

Sidgwick, Henry. The Methods of Ethics. 7th ed. London: MacMillan & Co. Ltd., 1907.

Simon, H. "A Behavioural Model of Rational Choice." Quarterly Journal of Economics 69 (1955): 99-118.

----------. "Theories of Decision Making in Economics and Behavioural Science." American Economic Review XLIX (1959): 253-83.

Slote, Michael. "Satisficing Consequentialism." Proceedings of the Aristotelian Society. Supplementary Volume LVIII (1984): 139-63.

----------. Common-Sense Morality and Consequentialism. Boston: Routledge and Kegan Paul, 1985.

----------. Beyond Optimizing: A Study of Rational Choice. Cambridge, Mass.: Harvard University Press, 1989.

Smart, J.J.C. "An Outline of a System of Utilitarian Ethics." In J.J.C. Smart and Bernard Williams, Utilitarianism: For and Against, pp. 1-74. Cambridge University Press, 1973.

Sorensen, Roy A. "Did the Intensity of My Preferences Double Last Night?" Philosophy of Science 53 (1986): 282-5.

Strasnick, Steven. "Social Choice and the Derivation of Rawls' Difference Principle." Journal of Philosophy 73 (1976): 85-99.

Tucker, Robert C., ed. The Marx-Engels Reader. 2nd. ed. New York: W.W. Norton & Company, 1978.

von Neumann, John and Morgenstern, Oskar. Theory of Games and Economic Behaviour. 2nd. ed. Princeton, N.J.: Princeton University Press, 1947.

Waldner, Ilmar. "The Empirical Meaningfulness of Interpersonal Utility Comparisons." Journal of Philosophy Vol. LXIX, No. 4 (1972): 87-103.

Weirich, Paul. "Utility Tempered with Equality." Nous Vol. XVII, No. 3 (1983): 423-39.

Williams, Bernard. "A Critique of Utilitarianism." In J.J.C. Smart and Bernard Williams, Utilitarianism: For and Against, pp. 75-150. Cambridge University Press, 1973.

Wong, Stanley. The Foundations of Paul Samuelson's Revealed Preference Theory. London: Routledge & Kegan Paul, 1978.