THE QUANTIFICATION OF 'HAPPINESS' IN UTILITARIANISM

THE QUANTIFICATION OF 'HAPPINESS' IN UTILITARIANISM

by

HOWARD JAMES SIMMONS, B.A., M.A.

A Thesis
Submitted to the School of Graduate Studies
in Partial Fulfilment of the Requirements
for the Degree
Doctor of Philosophy

McMaster University
September, 1986

DOCTOR OF PHILOSOPHY (1986) McMASTER UNIVERSITY
(Philosophy) Hamilton, Ontario

TITLE: The Quantification of 'Happiness' in Utilitarianism

AUTHOR: Howard James Simmons, B.A. (Oxford University)
M.A. (McMaster University)

SUPERVISOR: Professor J.E. Thomas

NUMBER OF PAGES: vii, 164

ABSTRACT

Utilitarianism is the theory that morality should be governed by the aim of maximizing satisfaction. But the concept of 'maximizing satisfaction' is a highly problematic one. This thesis attempts to resolve the difficulties.

After an introductory opening chapter, the main discussion begins with a defence of the hedonistic concept of 'satisfaction' in terms of pleasure and the absence of pain, in opposition to the currently more prevalent preference-oriented approach. An attempt is then made to explicate the concept of the 'intensity' of a pleasure or pain. An important consequence of the discussion is that pleasure and pain cannot in fact be put on the same metrical scale. Utilitarianism is thus seen really to have two separate components — a positive one, concerned with pleasure; and a negative one, concerned with suffering. These need to be clearly distinguished, although they will be isomorphic with respect to the solution of the maximization problem.

The discussion of this problem begins in Chapter Three. It is argued that conventional solutions — particularly Total Utilitarianism and Average Utilitarianism — are inadequate. The extreme view that 'numbers do not count' is also considered and rejected.

The fourth chapter contains my own proposal. According to the latter, the concepts of 'more pleasure' and 'less suffering' are not unitary in character, but are to be understood in terms of a multiplicity of principles of varying degrees of validity.

The concluding chapter discusses some outstanding difficulties and attempts to place the theory in a broader context.

Further technical elaboration of certain aspects of the theory is contained in two appendices to the thesis.

ACKNOWLEDGEMENTS

I would like to thank the members of my supervisory committee, Professors John Thomas, Douglas Odegard and David Hitchcock, for numerous pieces of helpful advice and criticism. I am also grateful to Professor Wilfrid Waluchow for discussing some of the issues with me. Finally, I would like to thank Chris Caskie for the speed and accuracy of her typing.

TABLE OF CONTENTS

Page

CHAPTER ONE: INTRODUCTION 1

Endnotes to Chapter One 15

CHAPTER TWO: THE ANALYSIS OF 'UTILITY' 17

1. Hare's Defence of Preference Utilitarianism 18

2. Harsanyi and the 'Equiprobability Model' . 27

3. The Concepts of Pleasure and Pain 35

4. Constructing a Hedonistic Metric 42

5. A Concession to the Sceptic 50

Endnotes to Chapter Two 53

CHAPTER THREE: THE EVALUATION OF SITUATIONS — CONVENTIONAL AND EXTREME APPROACHES 56

1. The 'Number-Intensity Conflict Case' 56

2. Brook and Schwimmer's Argument and a Response 59

3. Average Utilitarianism 67

4. 'Superlives' versus 'Equiprobability': Further Discussion 76

5. The Abandonment of Absolute Disvalues 80

Endnotes to Chapter Three 83

CHAPTER FOUR: CORRELATION AS THE BASIS OF UTILITARIAN SUPERIORITY 87

1. Synopsis of the Chapter 87

2. The First Two 'Perfect Principles' 89

3. A Third Correlative Principle 98

4. The Hierarchy of Perfect Principles and the General Principle NCP 105

5. The Collapse of the General Principle ... 110

6. The Imperfect Correlative Principles 117

Endnotes to Chapter Four 124

CHAPTER FIVE: CONCLUDING CHAPTER 127

1. The Balancing of <u>Prima Facie</u> Superiorities 127

2. Concluding Remarks 141

Endnotes to Chapter Five 143

APPENDIX A: PROOF OF THE EQUIVALENCE THESIS 145

Endnotes to Appendix A 157

APPENDIX B: THE NON-TRANSITIVITY OF CORRELATIVE SUPERIORITY 159

BIBLIOGRAPHY 163

CHAPTER ONE

INTRODUCTION

The purpose of this chapter is to develop a general theoretical structure into which the discussion of the specific problems with which this thesis is concerned can be fitted. Most of my efforts go towards defining, with as much rigour as is necessary for my purposes, certain key concepts, and explaining their significance.

First of all, we need the concept of a _situation_. The latter may be thought of as the entire contents of some region of space-time, or set of separate regions of space-time, in some possible world. For our purposes, the 'contents' of a region of space-time must be thought of as including any experiences that occur during that time for sentient beings located in that region (whether or not we actually believe that Mind-Brain Identity Theorists are correct in supposing that experiences literally have spatial locations).

I want now to introduce the concept of a _real preference_. I shall speak of an individual _i_ having, or not having, a real preference for some situation _X_ over some other situation _Y_ at time _t_. The idea that I want to capture here is that of a person's preferences with respect to the realization of different situations, *given* that he is adequately informed about their properties. This can be done by saying that *i* has a real preference for *X* over *Y* at *t* when there are conjunctions of properties *F* and *G* of *X* and *Y* respectively, such that if *i* knew at *t* that *X* had *F* and *Y* had *G*, he would as a direct result of this, prefer *X* to *Y*, whatever else he might also know about these situations. This definition takes for granted the notion of a 'preference', and I shall not further explain it, since this would take us into difficult problems inessential to the present work. The concept has, I think, enough intuitive clarity to justify its continued use here without any further explanation.

There are no limitations imposed on the kinds of properties that *F* and *G* might consist of. In particular, they are not limited to *internal* characteristics of the situations in question. For example, *i* might have a real preference for a situation in which some person who years previously committed some crime suffers over one in which he escapes suffering. The crime is not part of the actual situations themselves. But each situation has a relevant *external* property — the fact that it involves the person who committed this crime — which is the reason for the real preference. The admission of external properties also allows real preferences to be determined, partially or wholly, by instrumental considerations. *i* might have a real preference for *X* over *Y* because he prefers the *effects* of *X* to the effects of *Y*.

The principles underlying the determination of a person's real preferences at some time can be viewed in a manner reminiscent of the work of W. D. Ross. Two situations *X* and *Y* may be such that *i* would, if 'adequately' informed, find *X* preferable to *Y* *in a certain respect*, without finding it preferable *on the whole*, and therefore without having a real preference for *X* over *Y* in the sense in which I have defined that phrase. I shall then, partially *à la* Ross, speak of *X* being *prima facie preferable* to *Y* (for *i* at *t*).[1] I think it would be difficult to give an explicit definition of *prima facie* preference in terms of real preference. But the connection between the two concepts can be at least partially pinned down by the following 'axiom': if *X* is in some way *prima facie* preferable to *Y* for *i* at *t*, and *Y* is in no way *prima facie* preferable to *X* for *i* at *t*, then *i* has a real preference for *X* over *Y* at *t*. I shall call this the '*Ceteris Paribus* Axiom'. This axiom does not define what a real preference is. That is done by the formula on page 2. It might however be considered as providing an *implicit* definition of *prima facie* preferability in terms of the concept of a real preference.

When taken in combination with a complete list of the different forms of *prima facie* preferability valid for *i*, the *Ceteris Paribus* Axiom is sufficient to determine at least some of *i*'s real preferences. Whenever the properties of *X* and *Y* are such that *X* is *prima facie* preferable to *Y* in one of the ways listed, and *Y* is not *prima facie* preferable to *X* in any of the ways listed (which could always be determined, given the reasonable assumption that the list must be finite), then it logically follows that *i* has a real preference for *X* over *Y*. But because the converse of the *Ceteris Paribus* Axiom does *not* hold, not *all* of a person's real preferences can be determined in this way. The existence of a real preference for *X* over *Y* is compatible with the existence of conflicting *prima facie* preferabilities for *X* over *Y* *and* for *Y* over *X*. In such a case, the real preference is explained by the fact that the first set of preferabilities *outweighs* the second. A theory of the principles underlying a person's real preferences must involve both rules that determine the different kinds of *prima facie* preferability that are 'valid' for him and rules which determine their relative weights. However, the latter will not generally take the form of a straightforward *ranking*, so that certain kinds of *prima facie* preferability are seen always to outweigh certain others. Suppose, for example, that *i* has a *prima facie* preference for *X* over *Y* whenever *X* has less suffering than *Y* and also whenever *X* is less unjust than *Y*. And suppose that *A* has less suffering than *B*, but is also more unjust than *B*, but *A* and *B* do not differ in any other ways relevant to *i*'s preferences. It is unlikely that *i* would think that the suffering-criterion would always outweigh the injustice-criterion or vice-versa. Rather, he would probably think it important to consider how *much* less suffering there is in *A* than in *B* and how *much* more injustice. Thus a given kind of *prima facie* preferability does not usually have a universal significance or weight. Its weight varies from case to case according to the precise *way* in which it is realized.

In terms of these concepts, we can also give an idea of what it means to talk about the strength of a real preference. A real preference for *X* over *Y* is more or less strong to the extent that it exemplifies a weighty or significant way of realizing some *prima facie* preferability — i.e. one that is capable of overriding a wide range of ways of realizing *other* potentially competing *prima facie* preferabilities. Suppose, for example, that *i* has a very strong real preference for *A* over *B*, because *B* has vastly more suffering than *A*. And suppose *F* represents a description of the suffering in *A*, while *G* represents a description of the suffering in *B*. The real preference is based on the fact that if *i* knows that *A* has *F* and *B* has *G*, he will, as a direct result, prefer *A* to *B*, and this will be unaffected by any other knowledge he may have about *A* and *B*. Now the instantiation of *F* and *G* constitutes a particular way of realizing the relevant form of *prima facie* preferability — a form which I shall call 'N.H.U.S.' (for 'negative hedonistic utilitarian superiority'). And this way of realizing N.H.U.S.

is very significant or weighty, in the sense that it would override a very broad range of ways of realizing other sorts of *prima facie* preferability. And this is what makes *i*'s real preference for *A* over *B* a very strong one — its strength is a logical consequence of the particularly strong way in which N.H.U.S. is realized in this case. This is not to say of course that *nothing* could override this form of N.H.U.S.. There may be some pairs of property-conjunctions, such that if *i* believed that *A* and *B* had them, he would, despite his awareness of the vastly greater suffering in *B*, prefer *B* to *A*. (*Ex hypothesi*, however, *A* and *B* do *not* have them, and his belief would be mistaken. Otherwise, he could not be said to have a *real* preference for *A* — i.e. one that would be unaffected by whatever else he might know about *A* and *B*.)

Nothing has so far been done to throw any light on the concept of *morality*. To do this requires distinguishing between different *kinds* of *prima facie* principles. It would probably be very difficult to provide a strict set of necessary and sufficient conditions for a principle to count as 'moral'. But *one* important requirement, which has justifiably been given much attention by moral philosophers, is that of *universality*. This is that moral principles should not contain terms that refer to specific individuals. Thus if a theory for determining *i*'s real preferences at *t* includes a principle stating that any situation in which people are being considerate to him (*i*) is *prima facie* preferable to any situation lacking this feature, then this principle would not count as one of *i*'s *moral* principles. But a principle which stated that situations in which people are being considerate to other people generally are *prima facie* preferable to those lacking *this* feature *could* count as moral.[2]

Some moral *prima facie* preferabilities are also instrumental in character. An example would be if *A* were *prima facie* preferable to *B* for *i* because *A* would *lead* to people being more considerate to other people than *B*. Instrumental preferabilities generally are produced from other, more basic, preferabilities by means of a very important *prima facie* principle which applies, as a matter of logic, to *any* being capable of having attitudes to situations at all. This principle would run something like this: if *X* is *prima facie* preferable to *Y* and *X* and *Y* represent the total consequences of *X*' and *Y*' respectively, then *X*' is *prima facie* preferable to *Y*'. An instrumental preferability is *moral* when the more basic preferability from which it is derived is moral — i.e. determined by a principle which can be considered to be moral in character.

A *moral system* can be construed as a system of such principles, plus other principles which determine their relative weights (more explicitly — their relative weights in different manners of realization). Note that a moral system does not need principles which generate specifically *instru-*

mental kinds of moral preferability. For the latter are sufficiently explained by the more basic preferabilities from which they are derived plus the principle suggested above for producing instrumental preferabilities in general. I shall therefore say no more about instrumental preferabilities in this thesis.

A moral theory is the expression of the philosophical advocacy of a particular moral system, or narrow range of moral systems, plus a theory of how the acceptance of such systems by people will or would affect their behaviour. This latter element does require a separate mention, since the moral system only deals with the desirability and undesirability of situations, not with right and wrong action as such. Henceforth, I shall refer to the latter as the behavioural part of a moral theory, and to the former as the attitudinal part.

Utilitarian theory has an attitudinal component which advocates prima facie principles focusing on the general 'satisfaction' or 'dissatisfaction' within situations.[3] It also has a behavioural component which tries to explain the implications of the attitudinal theory for right and wrong action. The standard version of the behavioural component is that provided by Act Utilitarianism, which requires that each possible act be judged according to the value of the total situation that would result from it, given the circumstances. Act Utilitarianism has well-known difficulties. The attempt to avoid these difficulties led to the formulation of Rule Utilitarianism, which judges possible acts according to their conformity to rules, the rules themselves being judged according to the total situation that would result from everyone's conforming to them. In his recent book, *Utilitarianism and Co-operation*,[4] Donald Regan finds both approaches to be inadequate and attempts to supersede them by means of his theory of *Co-operative Utilitarianism*, according to which performance of the right action involves conformity to an overall *scheme* of action for a set of 'co-operating' agents, a scheme whose realization would produce the best overall consequences, given the behaviour of the non-co-operators. As Regan himself explains, his theory is an *objective* one, which does not make right action dependent on an agent's beliefs. The right action is the one that the rational utilitarian agent would perform if he were aware of *all* the relevant facts. Since such a condition is rarely, if ever, fulfilled, it might be argued that a *subjective* theory, which relativized rightness to the information actually available to the agent, would be more meaningful. Perhaps a subjective version of Regan's own theory could be produced. But I cannot pursue these questions here. For my main concern in this thesis is not in fact the behavioural component of moral theory, but rather its *attitudinal* component, that is, the theory of what makes certain situations morally preferable to certain others.

I must emphasize here that I am perfectly prepared to adopt a pluralistic approach to this matter. I see no reason to suppose that utilitarian criteria should be the only moral criteria for judging situations — that the utilitarian principle is the only valid _prima facie_ moral principle. Other sources of _prima facie_ moral preferability may well be recognized as acceptable, and they may on occasion conflict with the utilitarian principle.[5] (Some of them may — as in the example on page 2 — appeal to 'external' properties of situations, unlike the utilitarian principle which appeals only to 'internal' ones.) My concern here can be expressed in this question: to the extent that a moral system _is_ utilitarian, what form should it take? How should the utilitarian principle (for judging situations) be stated?

This thesis makes a number of suggestions with regard to this question. One is that the 'satisfactions' and 'dissatisfactions' which are held to be the pertinent feature of situations — their 'utilities' and 'disutilities' — should be understood in the traditional hedonistic way as states of happiness, and not in terms of the satisfaction of 'desires' or 'preferences'. However, it is also argued that, strictly speaking, there is no _single_ relation denoted by the phrase 'more happiness'. We must separate the question of whether one situation contains more pleasure than another from the question of whether it contains less suffering, since, as I shall argue in Chapter Two, it is not possible to put pleasure and suffering on the same metrical scale. Thus it is necessary to recognize two distinct *kinds* of *prima facie* preferability, one based on the concept of more pleasure, and the other based on the concept of less suffering, which, when they conflict, must be weighed against each other (as well as being considered in relation to any other non-utilitarian types of *prima facie* preferability that are recognized as valid). Furthermore, as I shall try to show in Chapter Four, the concepts of 'more pleasure' and 'less suffering' are themselves extremely complex, and in fact rather fuzzy, notions.

But this is yet to come. For the moment, I want to try to deal briefly with two issues that might be raised concerning the theoretical structure that I have described in this chapter. One is the fact that the structure is so obviously oriented towards *consequentialist* moralities, because it construes moral principles as determining preferences with respect to situations or states of affairs. Indeed such a structure does not even seem to *allow* for the possibility of deontological moralities, for which the fundamental concept is that of the rightness or wrongness of acts, rather than the goodness or badness of situations.

This objection does not seem to me to be well-founded. If one could give a system of rules which would determine *all* of a person's real preferences, this would surely constitute a complete representation of his attitudes. Even if his morality had a deontological component, its substantive content would have to be covered by such a system of rules. Deontologists must either be advocating certain kinds of attitudes rather than others, or they must be making conceptual points about the use of words like 'right', 'wrong', 'good', 'bad' and so on. The former is the _substantive_ content of a deontological ethic. And it must be represented in some way, either explicitly or implicitly, in a system of principles for determining real preferences. Of course, it would be perfectly correct to point out that attitudes do not _have_ to be represented in terms of preferences with respect to situations, conceived of as the contents of regions of space-time. But it is sufficient merely that they _can_ be represented in this way. To do so is convenient from the point of view of the discussion of utilitarianism, but it does not, as far as I can see, exclude the possibility of a deontological element in morality.

Another question that might be raised concerns methodology. I have said that in presenting a moral theory, a philosopher is _advocating_ certain moral systems in preference to others. But what is the basis of this advocacy? Why favour any one moral system over any other? It seems to me that to be acceptable a moral system should fulfill two basic conditions. Firstly, it should determine real preferences that are in accordance with our intuitions. It might be thought necessary to strengthen this condition by requiring that the intuitions be such that they would survive exposure to all the relevant facts. Now I do accept this requirement. If someone says 'There may be some facts which, if I knew them, would alter my attitude, but I would prefer not to know them', he cannot be engaging in serious moral deliberation or discussion — as a matter of definition. Notice, however, that this requirement is already built into our concept of a real preference. We can see from the definition on page 2 that if _i_ has a real preference for _X_ over _Y_, then, if we imagine _i_'s knowledge of _X_ and _Y_ progressively increased, starting from a position of total ignorance and continuing indefinitely (while everything else remains the same), we would eventually reach a point where _i_ would prefer _X_ to _Y_ (in the ordinary sense) and such that this preference would continue to obtain forever beyond it. It is a clear consequence of the definition of a real preference, then, that testing a moral system by comparing its consequences with our intuitions requires ensuring that those intuitions are as fully 'informed' as we can make them.

This first condition that a moral system should satisfy is the _substantive_ one. The second condition is that of _formal_ adequacy. A moral system might have principles whose consequences agree entirely with our 'corrected' intuitions, but fail to be as simple as it could be. For example, there might be a number of different principles whose work could be done by a single more general principle. In that case, the system will not satisfy the condition of formal adequacy, which thus further limits the range of acceptable moral systems. Note that the test of formal adequacy is only applied *after* a system has been seen to accord fully with the condition of substantive adequacy.

Of course, this methodology is in a very real sense 'subjective'. The most important criterion of adequacy is agreement with the particular intuitions or feelings of whoever is assessing the system. And although the intuitions do have to be capable of surviving exposure to all the relevant facts, the notion of 'relevance' used here is a very weak one. A fact is relevant for a given person if his knowing it *would* affect his attitude — we do not say that it is relevant when it *ought* to affect his attitude. There is still plenty of room for irresolvable disagreement, when facts which affect some people's attitudes do not affect others' — when people react differently to the same facts. My methodology is *subjective*, in the sense that it is not guaranteed to give the same results to all equally well-informed people who use it correctly. Methodologies of this kind fail to satisfy many philosophers who seek something more stringent. A very traditional course would be to strengthen the notion of an 'intuition'. Intuitions should not be mere subjective feelings, but ought rather to be cognitions of moral 'facts'. This is of course vulnerable to the objection that the notion of a moral fact seems irremediably obscure. Some kind of ethical non-cognitivism seems to make much more sense than the idea of there being literal moral 'truths'.[6] Thus although I recognize that the question is an extremely complex one, I am inclined to think that there is no reasonable alternative to settling for a subjective methodology. This will explain my strong emphasis on intuition in this thesis. At the same time, argument is an appropriate tool, not because those who reject good arguments are necessarily 'wrong' in an objective sense, but because arguments can be used in a persuasive manner to influence people's intuitions.

ENDNOTES

[1] See W. D. Ross, The Right and the Good (Oxford: Clarendon Press, 1930), especially pp. 19-20 and p. 138.

[2] I am not suggesting that universality is a *sufficient* condition for a principle's being moral, only that it is necessary. Another condition that one might impose is that a moral principle directly concern *persons* — or at least sentient beings.

[3] Thus by my definition, Moore's 'Ideal Utilitarianism' does not count as a genuine form of utilitarianism. It is certainly *consequentialist*. But since I am in a sense considering *all* moral systems to be basically consequentialist in character, this is not sufficient reason, from my point of view, for calling it 'utilitarian'.

[4] Donald Regan, Utilitarianism and Co-operation (Oxford: Clarendon Press; New York: Oxford University Press, 1980).

[5] In my final chapter, however, I explain how a person

can accept non-utilitarian moral criteria (in a non-instrumental way) and yet still count as a 'utilitarian' in some fairly all-encompassing sense. (see p. 139.)

[6] It is true that non-cognitivism does not always bring with it the rejection of 'objective' methodologies. R. M. Hare, a prescriptivist, believes that the peculiar features of moral language commit us to a precise and rigorous method of resolving moral disputes, a method which amounts in effect to Preference Utilitarianism. Although Hare himself is hostile to the subjective/objective dichotomy, it is pretty clear that this methodology is objective, according to the usage adopted here, that is, it is guaranteed to give the same results to all equally well-informed people who apply it correctly. However, while I more-or-less agree with Hare in his description of the conditions that discourse has to satisfy in order to count as 'moral', I do not believe that they lead to Preference Utilitarianism, or in fact to any other substantive moral position. (I try to establish the first of these two points in the next chapter.)

CHAPTER TWO

THE ANALYSIS OF 'UTILITY'

We have seen that a moral system that is partially utilitarian will be, to that extent, concerned with the relative values of 'states of affairs' or 'situations' as determined by the utilities and disutilities (individual units of 'satisfaction' and 'dissatisfaction') which they contain. There thus arise two distinct problems for the utilitarian. The first concerns the nature of the utilities and disutilities themselves — what exactly are they and how is their value to be measured? The second is the problem of how _situations_ — which may contain many different utilities and disutilities of different degrees of value or disvalue — are themselves to be compared with respect to value or disvalue. The first problem is the subject of the present chapter. (The second is dealt with in the succeeding two.) The first two sections constitute a critique of the view that utilities should be taken as _preference-satisfactions_. The rejection of this view leads, in Section Three, to a reconsideration of the more traditional hedonistic version of utilitarianism, beginning with an attempt to explicate the problematic notions of 'pleasurableness' and 'painfulness'. In the last two sections the question of the measurement of these properties is addressed. In Section Four it is argued that pains do admit of metrical comparison with other pains and pleasures with other pleasures, but in the final section such comparability is denied for the case where pleasures are compared with pains.

1. Hare's Defence of Preference Utilitarianism

In recent years the view that the aim of the utilitarian should be taken as the maximal satisfaction of people's preferences or desires has acquired much more popularity than the more traditional construal in terms of the maximization of happiness, the latter being a function of pleasure and the absence of pain. R. B. Brandt gives a helpful summary of the attractions in the preference-oriented approach:

> First, it allows that a wide variety of different states of affairs can be good — anything that can be wanted for itself. Second, it is thought easier to measure the strength of desires than to measure an amount of pleasure. Third, the desire theory may seem more democratic; it goes on the basis of what people in fact want, not on the basis of what will give them happiness — we are not to deny people what they want just because we think it will make them happier in the long run.[1]

But as defences of Preference Utilitarianism, these points are suggestive only. They do not provide any solid reason for pursuing the aim of bringing about, as much as possible, whatever people may happen to want. A solid reason is needed because the satisfaction of a preference *as such* is not something that we can immediately perceive to be intrinsically desirable, unlike the bringing about of a pleasurable experience, or the avoidance of a painful one. In this section and the following one, I consider two positive arguments that have been offered for the theory. Both maintain in effect that the aim of maximal preference-satisfaction is implied in the very nature of morality.

R. M. Hare believes that Preference Utilitarianism is forced on us by the logical properties of moral language. He maintains that every acceptable singular moral judgement concerning what should be done in some situation must require an action that would maximize the preference-satisfaction of all those affected by it. He constructs his argument for a simple case in which one person is contemplating performing some action that will harm or inconvenience another.[2] Let us call the first person '*A*', the second '*B*', and the contemplated action '*X*'. It is assumed that no other person would be significantly affected by whether or not *X* is done. Hare wishes to prove that *A* can only accept the singular moral judgement 'I ought to do *X*' (call this judgement '*s*') if his doing *X* would maximize the satisfaction of his *and* *B*'s preferences. This means that he can only accept it if his desire to do *X* is stronger than *B*'s desire not to experience the harm that *A*'s doing *X* will cause for *him*. (Note that when we talk about '*A*'s desire to do *X*', this must mean of course his selfish desire prior to moral deliberation. *After* moral deliberation his desire may be different.)

Hare's argument for this claim can be thought of as proceeding through these four stages:

(1) If *A* is well-informed and thus thoroughly understands what it would be like for *B*, in his particular situation, and with his particular psychological propensities, to experience the results of *X*, he will form a *prima facie* desire for *X* *not* to be done, were *he* (*A*) in precisely that situation, a *prima facie* desire equal in strength to *B*'s desire that it not be done in the *actual* case. (From now on, I shall refer to the actual case as '$\underline{C}_1$', and the hypothetical case in which *A* is in precisely the position that is occupied by *B* in $\underline{C}_1$ as '$\underline{C}_2$'.)

(2) The logic of moral language requires that any singular ought-judgement, and thus *s* (the judgement that *A* ought to do *X*), be both prescriptive and universalizable. This means that *A* can only sincerely accept *s* if it accords with his preferences *both* with regard to the actual case $\underline{C}_1$, *and* with regard to other possible cases that are exactly like $\underline{C}_1$ in their universal properties — and this includes $\underline{C}_2$.

(3) But if *A* is genuinely well-informed, these preferences will conflict. For we have seen in (1) that his being well-informed would give him a *prima facie* desire that *X* not be done in $\underline{C}_2$. But he also desires that, in $\underline{C}_1$, *X* *be* done. Thus *A*'s singular ought-judgement should (if *A* is rational) agree with the 'net' desire formed from these, which will simply be the stronger of the two.

(4) We have seen that the strength of the *prima facie* desire that *A* will have (if he is well-informed) that *X* not be done in $\underline{C}_2$ is equal to that of *B*'s desire that it not be done in $\underline{C}_1$. Thus, in virtue of (3), it follows that if *A* is well-informed and rational, he will only accept *s*- the judgement that he ought to do *X* in $\underline{C}_1$ — if his own desire to do it is greater than *B*'s desire that it not be done. And this is what needed to be proven.

In my criticism of Hare's argument, I shall try to establish two points:

(i) Hare does not make a strong enough case for the claim that a thorough understanding of *B*'s position would give *A* a *prima facie* preference that *X* not be done in $\underline{C}_2$ that was *equal* in strength to *B*'s preference that it not be done in $\underline{C_1}$.

(ii) Hare seems to misidentify the real implications of the requirement of universalizability. When these requirements are correctly understood, they do not generate his result.

There may be other points at which Hare's argument could be criticized besides these. But these appear damaging enough.

(i) Hare claims that if someone is not liking what is happening to him, then, if I have a thorough understanding of his situation, _I_ must desire (_prima facie_) to the same degree, that if _I_ were in the same situation, the same thing not happen to me. Clearly the fact that the desire is only _prima facie_ is important. If it were claimed to be absolute, it would follow that a thorough understanding of this person's situation was incompatible with consciously choosing to put oneself in the same situation, which is clearly false. But should we even accept the attenuated claim that a _prima facie_ desire must exist?

Hare defends this claim in two ways. The first is basically just an appeal to the unintuitiveness of denying the claim in a concrete case.

> Now consider our knowledge of what it is like to be _somebody else_ who is suffering (e.g. because his neck is being broken) . . . Suppose that I said 'Yes, I know just how you feel, but I don't mind in the least if somebody now does it to me': should I not show that I did not really know, or even believe, that it was like _that_? Would not my lack of knowledge or else my insincerity, be exposed if somebody said 'All right, if you don't

mind, let's try'?[3]

The trouble with this argument is that it attacks a thesis stronger than the denial of the thesis Hare is defending viz. that it would be possible for me to know exactly what it was like to be undergoing extreme suffering and yet not mind in the least if it happened to me.[4] But Hare's thesis is not merely that I would have to mind, but that I would have to mind to the same degree as the suffering person, to have an aversion equal in strength to the aversion felt by this person at the time of his suffering. And this hardly seems plausible. It would imply that no-one who knew what it was like to be undergoing some piece of extreme suffering — say the suffering of being tortured — could choose, for the sake of some ideal, to bring about that suffering for himself, unless his commitment to that ideal were strong enough to outweigh an aversion of equal strength to that which he would be feeling at the time of the suffering. But it is very unlikely that anyone's commitment is ever strong enough to outweigh so strong an aversion. This means that such a choice would never be made, which is surely false.

Hare's second way of defending his claim involves the suggestion that '. . . by calling some person "I", I express at least a considerably greater concern for the satisfaction of his preferences than for those of people whom I do not so designate'.[5] In other words, the claim can be defended by reference to the logic of the word 'I'.

But again, notice that the statement quoted, although very plausible, is also very weak, and does nothing to establish the claim that in identifying with a person one must take over their interests to the same degree. There does not seem, for example, to be any incoherence in my believing that I am going to suffer greatly two years hence — that it will be *I* who suffer — and yet not caring as much about it *now* as I will do at the time. Indeed, such a reduction in degree of concern owing to distance in time is the normal state of affairs.

The conclusion of this discussion is that the *most* we can accept is that *A* must have *some* desire that *X* not be done in $\underline{C}_2$, the hypothetical case in which he (*A*) is in the position actually occupied by *B* and will thus suffer the harm that *X* entails for anyone in that position. It has not been established that his desire must be as strong as *B*'s desire not to be treated in that way in the actual case. But clearly this is vital to the intended conclusion of the argument, for *B*'s desire not to be harmed must be given *equal* weight with *A*'s desire to do what will harm him. The former desire is represented *as* *A*'s desire that the harm not be done in $\underline{C}_2$. But we have no reason to deny that, in being represented in this way, its strength might be diluted.

(ii) According to Hare, *A*'s desire to do *X* in $\underline{C}_1$ is somehow in competition with his desire that it *not* be done in $\underline{C}_2$. But since these are preferences with respect to *different* possible situations, it is hard to see how they can conflict in any straightforward way. Hare argues that the requirement of universalizability demands that they both be taken into account. But the question of *how* they should be taken into account needs to be more carefully examined.

Universalizability requires that A can only sincerely accept that he ought to do X in C_1 if he accepts the universal moral principle that it ought to be done in *all* situations of the same general kind. Since C_2 is, by definition, of the same general kind, this means that it ought to be done in C_2. And, because of the prescriptivity of moral judgements, A cannot accept *this* unless he sincerely *prefers* that X be done in C_2. The rule operating here seems to be this: for the acceptance of a moral principle by someone to be sincere, its implications for every possible situation to which it applies must agree with that person's overall preference *with respect to that situation*. This is indeed highly plausible, but notice that it does not lead us into any process of *weighing* A's overall preference with respect to C_1 against his overall preference with respect to C_2, and observing which is the stronger. There will be no process of weighing, since these preferences must *agree* both with each other *and* with the moral principle, for the acceptance of that moral principle to be sincere. What *may* have to be subjected to a weighing-process is the *prima facie* preference which Hare believes A must have that X

not be done in $\underline{C}_2$. The purpose of this would be to determine _A_'s _overall_ preference with respect to $\underline{C}_2$. But this would obviously require us to weigh it against any _prima facie_ preference that _A_ may have that _X_ be done in $\underline{C}_2$, not — as Hare seems to demand — against _A_'s overall preference that _X_ be done in $\underline{C}_1$. And there is no reason to think that the former preference would coincide in strength with the latter.

There is another problem. When we say that _A_'s moral principle must agree with his overall preference with respect to each possible situation to which it applies, we must here mean, on pain of absurdity, his overall preference _subsequent_ to moral deliberation, not prior to it. But in that case, how can these preferences be used to place a constraint on what the moral principle should be? In order to know what the 'post-moral' overall preferences are, we need to know, not only the non-moral _prima facie_ preferences, but the moral ones also. And this means that we need to have _already_ answered the moral question. There thus seems to be a fundamental circularity here.

Thus Hare's attempt to show that _A_ is required to gear his singular moral judgement to whatever would be the result of weighing his desire to do _X_ in $\underline{C}_1$ with _B_'s desire that he _not_ do _X_ in $\underline{C}_1$ fails. Hare tried to show this by arguing that the judgement must agree with whatever would be the result of weighing _A_'s desire to do _X_ in $\underline{C}_1$ with his _prima facie_ desire that _X_ _not_ be done in $\underline{C}_2$, and that the latter was an adequate representation of _B_'s desire that _X_ not be done in $\underline{C}_1$. In (i) I showed that the second of these points is false, since there was no reason to believe that _A_'s _prima facie_ desire that _X_ not be done in C_2 would have to agree in strength with _B_'s desire that it not be done in $\underline{C}_1$. And in (ii) I showed that the _first_ point was also false. Thus Hare's argument for a Preference Utilitarian approach fails on (at least) these two crucial counts.

2. Harsanyi and the 'Equiprobability Model'

I now wish to consider another way of defending Preference Utilitarianism viz. that of John Harsanyi. It is somewhat similar to Hare's in its general approach in that it views the essence of morality as lying in an attitude of impartiality between the interests of all individuals. Harsanyi contrasts morality with self-interest. One who reasons simply according to the latter merely seeks to maximize the satisfaction of his _own_ preferences, while one who reasons from a moral standpoint has to take into account the interests of all individuals equally. But the details of Harsanyi's procedure are, as we shall see, somewhat different from those of Hare's.[6]

According to Harsanyi, a rational self-interested individual will, in choosing between a number of different alternative outcomes, choose the one whose utility for him is highest. (In cases where his utility in a given outcome is uncertain, he will seek to maximize his _expected_ utility, which is a function of each possible utility together with the probability of its realization.) The utility of a given outcome for a given individual is the degree to which his personal preferences are satisfied in it — or, more accurately, it is the degree of satisfaction of his 'true' preferences which are 'the preferences he _would_ have if he had all the relevant factual information, always reasoned with the greatest possible care, and were in a state of mind most conducive to rational choice.'[7]

Now Harsanyi is concerned with the problem of determining the correct 'social welfare function' i.e. the correct measure of a society's overall welfare. The social welfare function must be such that, if a rational individual who was reasoning _morally_, as opposed to merely self-interestedly, had to choose between a number of alternative possible societies, he would choose the one for which the value of the function was highest. Now moral reasoning imposes, by definition, a requirement of _impartiality_. The rational _moral_ deliberator must be concerned equally with the interests of _all_ the participants in the societies being considered. But, Harsanyi argues, reasoning impartially is equivalent to reasoning in a self-interested way under a certain kind of ignorance viz. an ignorance of precisely _who_ one is, and thus of what one's own particular interests are. Thus the moral deliberator will use a choice-prodecure which is effectively equivalent to that of someone who wishes to maximize his own personal gain, but who is ignorant of which particular position he will occupy in each possible society if that society is chosen — who believes indeed that he has an equal chance of occupying any given position. We have seen that a rational self-interested person chooses between alternative outcomes by determining in which one his individual utility is highest or, in situations of uncertainty, the one in which his individual expected utility is highest. In the present case, the different possible outcomes are the different societies. The situation is one of uncertainty because, since the chooser does not know what his position in any given society would be, he cannot assign a single definite utility to each outcome. But for each one, corresponding to each possible position that he might occupy, there is a possible utility, and the fact that he has an equal chance of occupying any given position generates a probability for each such possible utility (according to the number of positions that would generate that utility) which in turn generates an expected utility for that outcome. It is not difficult to see from this that the chooser will select the outcome — i.e. society — in which the mean utility per person is highest.[8] It is in this way then that Harsanyi justifies the identification of the social welfare function with mean utility, where 'utility' is understood in terms of degree of satisfaction of a person's 'true' preferences.

This theory produces an uncomfortable consequence. If the key feature in the evaluation of each outcome is the degree to which the preferences that people have in that outcome are satisfied, why not proceed by making preferences easier to satisfy rather than being enslaved by their presently very demanding preferences? Why shouldn't social planners try to cultivate in people an aversion to pleasure and a much greater tolerance of suffering than they otherwise would have? These preferences would not necessarily fail to qualify as 'true' preferences in Harsanyi's sense. They could survive exposure to rational consideration of the facts, since they would not have to be based on any false or unjustifiable beliefs. So it would appear that Harsanyi's version of the theory does not escape this familiar objection to Preference Utilitarianism.

This counterintuitive implication of Harsanyi's view should at least make us suspicious of his claim that it is implied by the concept of moral impartiality. And if we reexamine the basis for this claim we do indeed find it to be defective. The argument hinges on the suggestion that impartiality can be construed in terms of rational self-interest constrained by certain hypothetical conditions. But we find that the conditions as Harsanyi states them are not in fact strong enough to create genuine impartiality, while if they are appropriately strengthened, they turn out to render a self-interested decision logically impossible. Let us recall the situation. The rational self-interested chooser is required to select a society in which he himself is to be a participant. But he does not _know_ which participant he is to be in any given society which he might choose — he is stipulated to have an equal chance of being any given one. But note that this does not, in itself, guarantee the impartiality required. Suppose the chooser happens to be a successful entrepreneur. This may give him a bias towards capitalist societies which is not wholly dependent on a concern for his own position. He may well know that he is not guaranteed to be a successful entrepreneur in any capitalist society that he might choose, but this need not affect the existence of _some_ bias towards capitalism. Of course he does not know that he has this bias, since he does not know what any of his personal preferences are. He ranks certain possible societies above others without explicitly knowing why he has chosen one ranking rather than another. But there does not seem to be any incoherence in this. Nor could Harsanyi convincingly object that the preference for capitalism would not be a 'true' preference. There seems to be no reason to think that it would have to disappear as a result of rational consideration of the facts. (Of ccurse no _moral_ considerations can affect such a deliberation, on pain of circularity.)

Thus the conditions that Harsanyi lays down are not sufficient for full impartiality. They need to be strengthened to something like the following:

(1) The chooser does not know who he will be in the society he chooses.

(2) He believes that he has an equal chance of being any given individual and thus of having the particular personal preferences of that individual.

(3) His actual personal preferences have no effect on his choice.

Since the chooser is rational and self-interested he will (so the argument goes) seek to bring about a situation which maximizes the satisfaction of his personal preferences, whatever the latter may turn out to be. This means, given condition (2) and the reasoning described on page 29, that he will choose the society with the highest average level of personal-preference-satisfaction.

This argument is, however, mistaken. It depends on the assumption that rational self-interested agents have a desire to maximize the satisfaction of their personal preferences <u>whatever these may be</u>. But this is false, as can be seen from a consideration of the following case:

The Actual World is such that:

(i) An individual <u>A</u> prefers <u>p</u> to NOT-<u>p</u>.

(ii) NOT-<u>p</u> is the case.

(iii) <u>A</u> knows that NOT-<u>p</u> is the case.

A certain possible world *w* is such that:

(i') *A* prefers NOT-*p* to *p*, and more strongly than he prefers *p* to NOT-*p* in the Actual World.

(ii') *p* is the case.

(iii') *A* falsely believes NOT-*p* to be the case.

(iv') In all other relevant respects, *w* is identical to the Actual World.

It is clear that if he is rational and self-interested *A* will prefer *w* to the Actual World. This is so in virtue of the following facts:

(A) He would like *p* to be the case.

(B) *p* is not the case in the Actual World.

(C) *p* *is* the case in *w*.

(D) Although if *w* obtained he would *not* want *p* to be the case, he does not have to worry about *feeling* dissatisfied in *w*, since he would falsely believe *p* not to be the case.

(E) In all other relevant respects, *w* is the same as the Actual World.

But although *A* will prefer *w* to the Actual World, he actually has a *lower* level of preference-satisfaction in *w*. This is because his preference in *w* for NOT-*p* is stronger than his preference in the Actual World for *p*, and thus *w* accords less with what he wants *in* *w* than the actual state of affairs does with his *actual* wants. Furthermore, we can see that the fact that *A*'s preference in *w* for NOT-*p* is not satisfied in _w_ would not be for him even a _consideration_ against _w_ and in favour of the Actual World. His _actual_ preference for _p_ plus his recognition that he would not _feel_ dissatisfied in _w_, is all that counts. This shows that he cannot have any desire for maximal personal-preference-satisfaction _as such_. This is not to deny that rational self-interested agents do seek maximal satisfaction of their personal preferences. Rather it is to assert that their commitment to this aim is not logically independent of their commitment to the preferences _themselves_. What is fundamental is that they have the particular preferences that they do have, and in virtue of having _them_, seek _their_ maximal satisfaction. But what the above example shows is that they are not committed to, nor even mildly motivated by, the maximal satisfaction of their personal preferences _whatever these may happen to be_. The object of their desire is not, properly speaking, the maximal satisfaction of their personal preferences, but the maximal satisfaction of a _particular_ set of preferences, which is _in fact_ the set of their personal preferences.[9]

We can now see that a rational self-interested agent, under the special conditions that we are now assuming in order to guarantee impartiality, would not in fact be able to make a decision. Since he does not have any desire for maximal personal-preference-satisfaction _as such_, and since, by condition (3), he is prevented from being influenced by his actual personal preferences, there does not seem to be anything left which could control his deliberations. The conditions necessary to guarantee the kind of pure impartiality that Harsanyi is seeking seem in reality to be too tight to allow _any_ sort of decision, let alone a decision in favour of Preference Utilitarianism. (In Section Four, I shall give an account of a less pure, but more workable, notion of impartiality.)

3. The Concepts of Pleasure and Pain

We have found the case for Preference Utilitarianism, in the writings of two of its major proponents, to be seriously defective. We are therefore perhaps justified at this point in reconsidering the hedonistic version of utilitarianism, in which the aim is to maximize pleasure and minimize pain. The first problem which the hedonist faces is that of defining, or at least giving some adequate explication of, the terms 'pleasure' and 'pain'. Derek Parfit gives a good explanation of the central difficulty:

> _Narrow Hedonists_ assume, falsely, that pleasure and pain are two distinctive kinds of experience. Compare the pleasures of satisfying an intense thirst or lust, listening to music, solving an intellectual problem, reading a tragedy, and knowing that one's child is happy. These various experiences do not contain any distinctive common quality.[10]

Parallel remarks could be made concerning pains or types of suffering. The experience of having a headache is utterly unlike the experience of feeling afraid. It is very difficult to detect any single quality of 'unpleasant hedonic tone' — to use C. D. Broad's phrase[11] — which is supposedly shared by both.

What then is it about pleasurable experiences that makes them all pleasurable, and correspondingly for painful experiences? Here is Parfit's answer:

> What pains and pleasures have in common are their relations to our desires. On the use of 'pain' which has rational and moral significance, all pains are when experienced unwanted, and a pain is worse or greater the more it is unwanted. Similarly, all pleasures are when experienced wanted, and they are better or greater the more they are wanted. These are the claims of Preference-Hedonism. On this view, one of two experiences is more pleasant if it is preferred.[12]

But, as Parfit himself then goes on to admit, this is not in accordance with our ordinary uses of the words 'pleasure' and 'pain'. He cites an example from James Griffin. Apparently Freud, towards the end of his life, refused pain-killing drugs so that he could continue to think clearly. Freud was thus clearly in pain, but for the Preference-Hedonist his state of mind was a desirable one, because it was in accordance with his (Freud's) own wishes at the time he was experiencing it.[13] But this simply shows that Preference-Hedonism is not really a form of hedonism, as that is normally understood. The hedonist would prefer that Freud not be in pain, contrary to his own wishes. The concept of pleasure and pain that we are looking for must accord with this fact. But at least the first statement of the previous quotation does seem to hold true of pleasure and pain, as ordinarily understood. It does seem to be their relations to our desires that causes experiences to be classifiable under these labels. R. B. Brandt suggests the following:

> I think myself that what it is for some element of experience to be pleasant is just for it to be making the person, at the time, want to continue or repeat it, just for itself and not for extraneous reasons.[14]

Similarly the painfulness of an element of experience would be the fact of its creating, at the time, a desire (not based on extraneous reasons) for its own cessation. Now obviously the desire in question need only be prima facie; it does not have to be absolute. In the example described above, Freud had an absolute or 'net' desire to go on feeling pain. But he must, Brandt would maintain, at least have had a prima facie desire for the pain to cease, some basic urge which, if it had been stronger, would have prevailed over his desire to continue thinking and caused him to ask for an analgesic. This prima facie desire for the cessation of the experience, caused purely and simply by having that experience, and independent of extraneous reasons is, on this view, what the painfulness of the experience consisted in.

However, this analysis of pleasurableness and painfulness, in the precise form in which Brandt proposes it, runs into a serious difficulty. This difficulty arises from two facts. The first is that when a feeling involves pain or suffering, the fact that it does so is _intrinsic_ to it. It would not be qualitatively the _same_ feeling if it did not involve suffering. The second is that the causing of a desire of the relevant kind cannot be an intrinsic feature of a painful feeling. These two facts jointly entail that the causing of the relevant desire cannot be _identified_ with the painfulness of the feeling. I shall now try to explain why I believe these two facts to obtain.

The first is, I think, easier to appreciate than the second. The claim is that the painfulness of a feeling is an intrinsic or essential feature of it, in the sense that any feeling that was not painful would have to be a _different_ feeling i.e. qualitatively distinct. Consider a toothache. It would clearly not be the feeling that it is if it were not painful. The same is true of more 'emotional' kinds of suffering. A feeling of fear cannot have its 'unpleasantness' removed without changing its intrinsic _content_. Brandt maintains that with regard to some pains, '. . . there is surgery which apparently leaves the sensation intact but reduces the unpleasantness of it to a rather mild level, so that it appears we must distinguish between the intensity of pain and the unpleasantness of it . . .'.[15] But if it is really true that patients who have undergone such surgery report having the very same sensations as before, while finding them less painful, I do not think that their statements can be taken literally. It is logically inconceivable that one could reduce the painfulness of a sensation without changing its qualitative content (where 'qualitative content' is understood to include intensity). Its very existence as the precise kind of sensation it is entails its painfulness.

The existence of a desire for the cessation of a feeling or sensation cannot, in the same way, be necessarily entailed by the content of that feeling or sensation. Such a desire must be either occurrent or dispositional in nature. Suppose, first, that it is occurrent. Then it consists simply in a feeling of wishing that the sensation would cease. Now note firstly that such an occurrent wishing obviously cannot be identical with the very sensation for whose cessation it _is_ a wishing. It must be distinct from it. But in that case, we can invoke Hume's observation that nothing can be logically deduced from the nature of one object concerning the existence or nature of other distinct objects.[16] The existence of the painful sensation can only _contingently_ be accompanied by an occurrent wish that it should cease. Suppose, on the other hand, that the desire is dispositional in nature. That is, it consists in a propensity to manifest, under the right circumstances, either an occurrent wish of the kind just considered, or certain kinds of behaviour designed to bring the sensation to an end. Such a propensity or disposition will consist in the truth of (contingent) conditional statements, with consequents asserting something other than the existence of the painful sensation. And for essentially the same reason as before, the existence of the sensation cannot logically entail the truth of such a conditional.[17] So whether the desire for the cessation of the pain be occurrent or dispositional in nature, it will always have to be extrinsic to that pain.

This must not of course be taken as a denial that a painful feeling is nearly always accompanied by such a desire. It may even be psychologically or factually impossible that this not be so. The point is merely that this desire is not logically guaranteed by the qualitative content of the feeling itself. And since the painfulness of the sensation is logically guaranteed by that content, this painfulness or 'unpleasantness' cannot be identified with the fact that such a desire is caused to exist.

But then this brings us straight back to the problem of trying to say what painfulness should be identified with. The fact that a feeling's painfulness is logically inseparable from its content recalls the suggestion that it is a distinctive quality of sensations — 'unpleasant hedonic tone'. But we have already noted the implausibility of this suggestion. There just does not seem to be any such distinctive quality which is shared by all painful sensations. The painfulness of a sensation seems somehow to permeate its entire content, rather than being a separable element in respect of which it can resemble sensations of otherwise distinct kinds. There is however a way in which we can retain the logical link between the content of a feeling and its painfulness without invoking 'unpleasant hedonic tone'. This method involves a subtle modification to Brandt's proposal. To be having a painful sensation is not _essentially_ to be having a sensation which is creating a desire that it should cease. But it is to be experiencing one of a large class of _types_ of sensation, such that the condition for a given type to belong to this class is that instances of it _in fact_ generally do give rise to such desires when they occur. This is not a distinction without a difference, as it might at first sight appear to be. The condition for a sensation to be painful is for it to be a token of one of a particular _class_ of types. The essence of that class is its extension, as is the essence of any class. Nothing more is required for a sensation to be painful than for it to be one of _those_ types. To explain which these are we must make reference to the fact that tokens of them do generally give rise to desires for their cessation. But this does not imply that painfulness itself necessarily entails the existence of such a desire. What it necessarily entails is merely the exemplification of one of the types in question, with the distinctive qualitative character of _that type_.

Parallel to the above analysis, we may say that a sensation is pleasurable when it exemplifies any of a large class of types of feeling which have the common property of generally being exemplified by tokens which simultaneously create a (non-extraneously based) desire for their continuation.

This account accords well with the hedonist's desire to say that pleasure is, as such, always good, and pain, as such, always bad. His thinking that pleasure is good, for example, is (at least approximately) the fact of his having a pro-attitude towards those sensation-types which, when exemplified, tend to create a desire for their continuation. But the fact of their creating such a desire, although no doubt causally connected with the existence of his pro-attitude, is not an essential part of the object of that attitude. Clearly, liking all things that have a certain property does not necessarily amount to liking them for that property. It is the types themselves, with their distinctive contents, which are the objects of his approval. Contemplation of these contents is enough in itself to cause him to seek experiences exemplifying them. He need not consider, as such, the desires which they will create at the time of their occurrence.

4. Constructing a Hedonistic Metric

One of the most notorious problems with which Hedonistic Utilitarianism is confronted is that of providing a theoretical basis for the notion that experiences have determinate degrees of pleasurableness or painfulness. It was pointed out in the quotation from Brandt in Section One that one of the major attractions of Preference Utilitarianism lies in its avoidance of this problem. Clearly then a return to Hedonistic Utilitarianism requires at least a reasonably plausible solution to the problem. In this section I shall talk almost entirely about pains, though what I say applies, with an appropriate shift in terminology, to pleasures as well.

The idea of measuring painfulness provokes the most scepticism in cases where we allow the type of pain to vary e.g. when we compare a 'physical' pain like a headache with an 'emotional' pain like a feeling of grief. The more dissimilar the types, the more difficult it is to believe in metrical comparability. How can we counter this scepticism?

One thing we cannot do is simply identify the greater painfulness of an experience with its having associated with it a stronger desire for its own cessation. For we have seen that the existence of such a desire is not logically essential to the painfulness of the experience as such. However, just as the contingent existence of such desires enabled us to explicate the notion of the painfulness of an experience in an indirect way, so it might be suggested that we could adopt a similar indirect explication of an experience's degree of painfulness, using the contingent connection between experiences and strengths of desire. I am not going to claim that such an approach cannot be made workable. There are however certain difficulties in its development which can be avoided by adopting a somewhat different line of attack, which I shall now endeavour to explain.[18]

This approach requires a closer look at the concept of intensity as this applies to mental phenomena generally, and not just to those which can be described as 'pleasurable' or 'painful'. 'Intensity' is probably not a univocal term, but according to one very important use of it, an experience is rendered more intense when it is, so to speak, 'magnified'. The most obvious example lies in the area of auditory experience. When two auditory experiences differ only in respect of their phenomenal volume, they share the same basic character, but one exemplifies that character to a greater degree than the other. And it is clear that a cardinal measure is applicable here. We might say, for example, that one of the sound-experiences had three times as much phenomenal volume as the other, meaning that the first exemplified three times as much as the second the basic character common to both. It is true that such judgements are often very approximate, but they are always capable in principle of being rendered more exact. Now what is true of auditory experiences seems in fact to be true of experiences generally. They are all capable of being assessed in terms of intensity, understood in this generic sense, a sense which is not conceptually tied to any particular kind of experience.

Now we have already seen in the previous section that the painfulness of a painful feeling cannot be separated from its intrinsic qualitative character. To double the intensity of such a feeling is to double the degree of realization of that character. It is thus *ipso facto* to double its painfulness. To assert that a painful sensation had doubled in intensity, but deny that the degree of painfulness had doubled would be to affirm an incoherent separability of a sensation's painfulness from the sensation itself. The degree of painfulness of a painful feeling is thus — on a first approximation — the intensity — in the generic sense — of that sensation.

But so far this approach only enables us to deal with the least difficult aspect of our problem — the comparison of pains which are all of precisely the same type. Since degree of intensity is clearly determinate in such a case, so must degree of painfulness be. But when we come to compare sensations of *different* types, it may be doubted that we can attach any meaning to the assertion that one is exactly as intense as the other. Although it may be one *concept* of intensity which is applicable to all of them, this does not of itself guarantee that there is one single scale on which all their intensities can be measured.

Consider any two pain-types $\underline{F}$ and $\underline{F}'$. Once we have found just *one* pair of intensities $\underline{i}$ and $\underline{i}'$ from each type respectively which can be equated with respect to degree of painfulness (i.e. such that *i'* is the *unique* equivalent in *F'* of *i* in *F* and vice-versa), we will know for *any* intensity *j* within *F*, what its unique equivalent in *F'* is (if it has one). For *j* must be *i* multiplied by some factor. Its equivalent in *F'* will simply be *i'* multiplied by the *same* factor. Since a change in intensity by the very *same* factor is simply a 'magnification' (or 'diminution' where the factor is fractional) of each pain to the very same degree, it must preserve the equivalence. The problem remains however of forming the *initial* link between the two scales represented by *F* and *F'*.

To do this, I shall define a concept of the *minimally objectionable intensity* within a given pain-type *F*. Imagine a person, who is understood to be motivated solely by the desire to avoid pain, subjected to gradually (and uniformly) increasing intensities within *F*. There is one particular action which he can perform in order to put an end to the process, say pressing a button. The minimally objectionable intensity is the *first* intensity which motivates him to press the button. We can then say that all pains which are of the minimally objectionable intensity of their types are equally painful. And then, in view of the reasoning of the previous paragraph, *all* the pain-types can be collapsed into one scale. For each point on the scale, there will be some *n* such that every pain at that point is *n* times as intense as the minimally objectionable pain of its particular type. A complete ordering with respect to degree of painfulness is achieved. And even a cardinal measure of painfulness is provided by the cardinal measure of intensities.[19]

There seems however to be a problem in this proposal, in that it may be the case that the minimally objectionable intensity for pains of a given type is not the same for all people at all times. The result of the empirical test described above may not be the same for all subjects at all times. This is of course essentially the so-called problem of the 'interpersonal comparison of utilities' in its hedonistic version. It might be suggested that the degree of painfulness of a pain of type _F_ experienced by a person _A_ at time _t_ is simply the degree to which it is more or less intense than the minimally objectionable intensity within _F_ _for A at t_, understood in terms of the result that the test would generate if it _were_ to be done on _A_ at _t_. The degree of painfulness of a feeling is thus relativized to the person experiencing it and the time at which he experiences it. This would take us back into a partially preference-oriented approach, similar to the theory described by Parfit on page 36. The individual is, to some extent at least, his own authority on the desirability of his state of mind at any given time. But although attractive, this approach will not work, as far as the Hedonistic Utilitarian is concerned. For if indeed it were to turn out that the test for minimal objectionableness gave different results for different people at different times then we would sometimes end up assigning different degrees of painfulness to the same pain as experienced by different people or by the same person on different occasions. And the Hedonistic Utilitarian does not want to say that the very same state of mind could, in different instantiations, have different values or disvalues.

The correct course to adopt is, it seems to me, to relativize minimal objectionableness to whichever individual utilitarian chooser happens to be making the moral decisions and the time at which he is making them. And no better justification for this is needed than the fact that they are his decisions. For there to be more pleasure and less pain in the world is, in itself, just a personal preference he has, even though it happens to be shared by many others.[20] The character of the situation is thus not fundamentally altered by allowing his personal preferences to dictate to some extent how pleasurableness and painfulness are to be measured. Thus at the ideal level of theory, the content of one moral agent's utilitarianism may be slightly different from that of another. But in the practical application of utilitarianism, where exact measurements of pleasure and pain are not possible anyway, these differences are unlikely to be greatly felt.

The reader may perhaps now have some indication of the kind of account of utilitarian impartiality I would substitute for the account given by Harsanyi and rejected in Section Two. A Hedonistic Utilitarian chooser (or any chooser to the extent that he is a Hedonistic Utilitarian) is impartial in two senses: first, he does not let the 'ownership' of a given state of mind affect his assessment of its intrinsic value — only its content counts. Secondly, in determining what that content *is*, he is not irrationally affected by a consideration of the content that would characterize his *own* state of mind were he in the same external situation. In assessing the impact of events on others, it is what it actually feels like for *them* that matters, not what it *would* feel like for the moral deliberator were he in their position. However, once he has decided (to the best of his ability) what it *does* feel like for them, he has to decide whether, and, to what extent, *he* wishes there to be people feeling like that. And he will do this in effect by determining how each feeling compares with his minimally desirable or minimally objectionable feeling of the same type. ('Minimally desirable intensity' applies of course to pleasures and is defined in a way parallel to 'minimally objectionable intensity'.) Thus the kind of impartiality embodied in Hedonistic Utilitarianism is not so pure as to rule out all forms of individual bias. But such purity would arguably render decision-making impossible anyway.

5. A Concession to the Sceptic

We have discovered a reasonably satisfactory way of construing the assertion that some feeling has a certain degree of painfulness or pleasurableness. This provides us with a theoretical criterion for choosing between pains and between pleasures. It would be natural to suppose that we could also use these notions to make decisions in situations where we must choose between having a certain pleasure and *avoiding* a certain pain. Unfortunately this does not seem to be the case. Although it does not seem unreasonable to define 'Pleasure *a* is as pleasurable as pain *b* is painful' as 'For some *n*, *a* is *n* times as intense as the minimally desirable pleasure of *a*'s type and *b* is *n* times as intense as the minimally objectionable pain of *b*'s type', one can plausibly argue that a hedonistically motivated person does not have to be indifferent between having one sensation and avoiding another which is as painful as the first is pleasurable.[21] To see this, compare the following two pairs of choices:

(1A) A hedonistically motivated person may have either of the two minimally objectionable painful feelings *a* and *b*.

(1B) The same person may have either of two painful feelings *a*' and *b*', such that, for some very large *n*, *a*' is of the same type as *a*, but *n* times as intense, and *b*' is of the same type as _b_, but again _n_ times as intense.

(2A) The same person may have EITHER minimally desirable pleasure _c_ plus minimally objectionable pain _d_ OR hedonically neutral feelings.

(2B) The same person may have EITHER pleasure _c_' plus pain d' such that, for some very large _n_, _c_' is of the same type as _c_, but _n_ times as intense, and similarly _d_' is of the same type as _d_, but again _n_ times as intense OR hedonically neutral feelings.

It seems clear that in cases (1A) and (2A) this person should be indifferent between the two options. And it is almost as clear that he should be indifferent in (1B) also. (This indeed is essential to the findings of the previous section.) For the choice in (1B) is simply the _same_ choice as in (1A) but at a higher level, so to speak. Each alternative is subjected to precisely the same change, and so the increases cancel one another out. But this is not so when we consider the relation between (2A) and (2B). Although it is true that one of the alternatives in (2B) can be obtained from one of the alternatives in (2A) by making precisely the same change to its two _components_, it is _not_ the case that (2B) can be obtained from (2A) by altering both _alternatives_ in the same way. Indeed one of the alternatives remains unchanged. Thus in moving from (2A) to (2B), we do not get the same sort of cancelling-out effect as we do when we move from (1A) to (1B). Although our chooser must indeed react in the same way when faced with (1A) or (1B) (viz. indifference between the two alternatives), he does not have to have the same reaction in (2B) as he does in (2A). Even though he is indifferent between the alternatives in (2A), he does not have to be indifferent between the alternatives in (2B). Indeed, if his view of the relative importance of avoiding extreme suffering and experiencing ecstatic pleasure is like that of most of us, he will prefer the alternative of hedonically neutral feelings.

Thus equal intervals in the positive and negative ranges do not always have the same importance. Now this does not of itself rule out the following possibility: that we might have a metrical scale in which the units were essentially units of *value*, having different positive and negative 'interpretations'. Thus for some *n* and *m*, one unit would represent a difference of *n* units of pleasure and *m* units of pain. But even this course is not viable. Consider the most agonizing pain imaginable. It must represent some number of units in the scale, say 50. But in that case its avoidance must be of equivalent value to the experiencing of 50 x *n* units of pleasure. But it is very unlikely that *any* pleasurable experience could be as worth having as this pain is worth avoiding. It seems therefore that we cannot accept the idea that there is — even for a single utilitarian chooser — one scale on which both pleasure and suffering can be placed. Thus when we come to discuss the relative value of entire 'situations', we should not — in the traditional manner — conceive of them as consisting of positive and negative units which can be 'processed' together, so to speak. Rather, we should consider the positive and negative aspects separately. How the positive and the negative relate to each other in the overall utilitarian scheme will be discussed in the last chapter.

ENDNOTES

[1] R. B. Brandt, "Problems of Contemporary Utilitarianism: Real and Alleged", in Ethical Theory, ed. by Norman E. Bowie (Indianapolis/Cambridge: Hackett Publishing Company, 1983), p. 88. However, Brandt is himself somewhat critical of Preference Utilitarianism. See his criticisms in A Theory of the Good and the Right (Oxford: Clarendon Press, 1979), pp. 247-253. Another writer who is sceptical of the theory is Lars Bergström ("Interpersonal Utility Comparisons", Grazer Philosophische Studien, 16/17 (1982), p. 291).

[2] R. M. Hare, Moral Thinking: Its Levels, Method and Point (Oxford: Clarendon Press, 1981), pp. 107-111. It is important to note that this thesis of Hare's that we are considering here applies only at what he refers to as the 'critical' level of moral thinking. 'Intuitive' thinking is different, using simple common sense prima facie principles, the choice of which is itself, however, governed by critical thinking. See particularly Sec. 2.5 of Chapter One, pp. 39-41, and the bottom paragraph of p. 113.

[3] Ibid., p. 94.

[4] In fact, it is one of the consequences of the findings of the third section of the present chapter that this is logically possible. For I argue there that even experiencing suffering does not logically entail any particular attitude to that suffering. So contemplating it certainly should not. But I do not have to insist on this point in order to undermine Hare's argument.

[5] Ibid., p. 98.

[6] Harsanyi's theory is presented in his "Morality and the Theory of Rational Behaviour", in *Utilitarianism and Beyond*, ed. by Amartya Sen and Bernard Williams (Cambridge: Cambridge University Press, 1982. Paris: Editions de la Maison des Sciences de l'Homme, 1982), pp. 39-62. See particularly, for the purposes of the present discussion, pp. 44-48 and pp. 54-56. On p. 54, Harsanyi argues briefly against Hedonistic Utilitarianism. But his argument consists simply in pointing out that Psychological Hedonism — the theory that people only ever desire pleasure or the avoidance of pain — is highly implausible. Since *Ethical* Hedonism — or theories that incorporate an Ethical Hedonistic component — need not entail or presuppose Psychological Hedonism, this is irrelevant.

[7] *Ibid*., p. 55. There is obviously some resemblance between this notion of a 'true' preference and the concept — explained in Chapter One — of a 'real preference'. Of course, it hardly needs to be pointed out that there is no incompatibility between using 'preferences' to explain what a moral system *is* (as in that chapter) and rejecting their use as the key notion in the substantive *content* of a moral system.

[8] Where n is the total number of people in the society, $U_1 \ldots U_r$ are all the possible utility-values, and $n_1 \ldots n_r$ are the numbers of positions corresponding to each of these utility-values, the expected utility is $\sum_{i=1}^{r} \frac{n_i U_i}{n}$, which is of course equivalent to the mean utility in the society. (Harsanyi has a slightly different, but equivalent, formal presentation of this on pp. 45-46.)

[9] For the sake of simplicity, I have here neglected the fact that Harsanyi talks in terms of 'true' preferences and not just preferences of any kind. To re-state the point in these terms, one could say that what Harsanyi *needs* is the proposition that rational self-interested individuals 'truly' desire the maximal satisfaction of their 'true' personal preferences, whatever these may be, whereas what is the *case* is that they 'truly' desire the maximal satisfaction of a particular set of preferences, which are *in fact* their 'true' personal preferences.

[10] Derek Parfit, *Reasons and Persons* (Oxford: Clarendon Press, 1984), p. 493.

[11] C. D. Broad, *Five Types of Ethical Theory* (London: Routledge & Kegan Paul Ltd., 1930), p. 230.

[12] Parfit, *Reasons and Persons*, p. 493.

[13] _Ibid_., p. 494.

[14] Brandt, "Problems of Contemporary Utilitarianism", p. 88.

[15] Brandt, _A Theory of the Good and the Right_, p. 37.

[16] It can be questioned whether Hume's thesis really holds in a fully general form. It can plausibly be maintained, for example, that the essence of _physical_ objects lies in their causal powers to affect other objects in certain ways. But with regard to mental objects at least, the thesis appears to be acceptable.

[17] When _q_ and _r_ are logically independent, and $\underline{p} \supset \underline{q}$ is contingent, the latter cannot be entailed by _r_.

[18] The main problem is that, at first sight at least, the precise determinateness of 'strengths of desire' is no clearer than that of painfulness itself. But I am not going to claim that an adequate account could not be successfully worked out. Brandt's behaviouristic approach in _A Theory of the Good and the Right_ may well be the answer. On the whole, I prefer a 'phenomenological' approach to a behaviouristic one, as being more intuitive, although I do have recourse to the behavioural in my explanation of the 'minimally objectionable intensity' on p. 46.

[19] In future, when I speak of the 'intensity' of pleasures and pains, I intend it to be understood in accordance with this conception (relativized, as I am about to explain, to the particular utilitarian decision-maker and the time of his decision). _This_ use of 'intensity' is not itself a generic use applicable to experiences of all kinds, since it involves the concept of 'minimal objectionableness', which is applicable only to painful experiences. But the explication of it does make use of the generic concept.

[20] The reader may not be surprised that I would say this, in view of my acceptance, in Chapter One, of a 'subjective' methodology. (See p. 14.)

[21] We assume of course that the experiences are of equal duration. The problems arising from varying durations are discussed in later chapters.

CHAPTER THREE

THE EVALUATION OF SITUATIONS — CONVENTIONAL AND EXTREME APPROACHES

I turn now to the second of the two major problems identified at the beginning of the previous chapter — the question of the comparative value of situations, which may be characterized by a multiplicity of different experiences with different hedonic intensities. Now it has already been pointed out that we cannot talk of utilitarian superiority *simpliciter*; we must separate *positive* utilitarian superiority — which is a matter of the maximization of pleasure — from *negative* utilitarian superiority — which is a matter of the minimization of pain or suffering. Since it is usually considered a more pressing moral concern, I have chosen to gear my discussion towards the latter rather than the former. But the treatment of the former ought to take a parallel course.

1. The 'Number-Intensity Conflict Case'

The most familiar utilitarian principle for deciding between alternative situations is the Principle of Total Utility, according to which the best situation is the one with the highest sum of utilities minus disutilities. Adapting the principle to the present discussion, we can divide the suffering of each person in the situations being compared into periods of equal duration and constant intensity of suffering[1], assigning a positive integer to represent each such intensity, and then state the principle by saying that one situation is negatively utilitarianly superior to another when its total sum of intensities is lower than that of the second. Is this a viable principle?

Consider the following: situation _A_ consists entirely of one thousand people each suffering a minor discomfort (of equal duration) whose intensity we shall represent by the number 1. Situation _B_, in contrast, consists entirely of just one person who is experiencing agonizing suffering (of the same duration as the discomforts in _A_) whose intensity we shall represent by the number 20. Which of the two situations, if either, is the better one?

One feature of the choice is the fact that the aim of minimizing intensity of suffering is in conflict with the aim of minimizing the _number_ of people who suffer. If we think that the former aim should take priority, we will choose situation _A_; if the latter, situation _B_. For this reason, I shall refer to this case as the 'Number-Intensity Conflict Case' (or for short, 'N.I.C.C.'). On the other hand, it seems intuitively clear how the conflict should be resolved. Situation _A_ appears to be much better than situation _B_.[2]

But of course this is not the answer given by the Principle of Total Utility. According to the latter, the total disvalue of _A_ is 1000, while that of _B_ is only 20. This would entail that _B_ is a much better situation than _A_.

Some people react to this sort of problem as if it were an instance of the well-known difficulty of accommodating utilitarianism to the notion of _justice_. These people would say that in allowing one person to suffer agony in order to spare a thousand people a mere discomfort, we would be unjustly exploiting that one person for the sake of the majority. Others see the issue as a conflict between utility and equality.[3] People who adopt either of these stances may not question the thesis that happiness is maximized in situation _B_ rather than in situation _A_, that there is more happiness, or rather less suffering, overall in _B_. They may accept that thesis, but insist despite that, that _A_ is preferable. However, it is possible to take a different view viz. that _A_ is preferable to _B_, because happiness is maximized in _A_, not in _B_, contrary to the position of the Total Utilitarian. This is the position that I want to explore here.

Indeed the suggestion that there is more happiness — or rather less suffering — in situation _A_ does not seem to be particularly implausible. It is true that there are a vast number of suffering people in _A_. But in the case of each one of these people, the _degree_ of suffering is only very minor. It is hard to believe that a minor discomfort, by simply _recurring_ in a vast number of people, can 'add up' to something as bad as agony. This suggests that the Total Utility Principle generates its implausible result by putting too much weight on the aim of reducing the *number* of suffering people as against the aim of reducing the *intensity* of suffering.[4] But what *should* the relative weights of these two aims be? One very radical suggestion is that numbers as such should not count at all. Only intensity should count. On this view, happiness is not 'interpersonally additive'. This thesis need not be stated in a purely hedonistic form. It may be asserted that goodness generally is not 'interpersonally additive', that producing good for more people does not, in itself, produce *more goodness* overall. I want now to consider a recent defence of this view.

2. Brook and Schwimmer's Argument and a Response

Richard Brook and Seymour Schwimmer have denied that the Good is interpersonally additive.[5] They maintain, for example, that if one course of action enables us to save four lives while another enables us to save only one, this is not in itself a good reason for choosing the first course of action. More relevantly for our purposes, they maintain that more people suffering does not necessarily make things worse, even *prima facie*. That is, it is not a good reason for choosing one course of action over another that it would lead to fewer people suffering. Their argument hinges on the observation that the divisibility of something into definite units does not of itself guarantee that it is additive over those units. '. . . Individuation, though necessary, is not sufficient for addition'.[6] They give the example of two groups of aircraft, one more numerous than the other, but all flying in formation at the same speed. Clearly there is no more <u>speed</u> involved in the more numerous group than in the less numerous one. Another example involves a comparison between one group of people, all of the same height, and a larger group, also of the same height as those in the first group. Again, there is no more <u>height</u> involved in the second group than in the first — unless there is some special circumstance, such as that the people in both groups are standing on one another's shoulders. In both examples, we can identify definite units of individuation — aircraft in the first case, and people in the second — but it does not necessarily follow that the property in question is additive over these units. Whether addition does or does not obtain depends on the special circumstances of the case. Further reason, beyond the mere existence of units, is needed to demonstrate the possibility of addition. Thus from the fact that there are definite units involved in doing good or relieving suffering — viz. people — it does not follow that good or suffering are additive over these units. Thus there is no reason to think that we are doing more good when we save the lives of more people instead of fewer. How then should we decide in such cases? Brook and Schwimmer believe that our impartial moral concern is adequately expressed by giving each person involved an _equal chance_ of being saved. (This might mean, in the case of a choice between only two groups, tossing a coin.)

It is of course vital to note that Brook and Schwimmer are talking here about _disjoint_ groups of people. If one group is a subset of the other — that is, if we have a choice between helping certain people and helping these _and_ certain others, we should do the latter. This is what Brook and Schwimmer call the 'Pareto Exception'. The admission of the Pareto Exception is not, they maintain, inconsistent with their belief that numbers _as such_ do not count, since the reason for helping the larger group in this case would not be that in doing so one would be helping a larger number of people, but rather that in _not_ doing so, one would be failing to help certain people whom one _could_ help at no extra cost.

What, if anything, can be said against Brook and Schwimmer's argument? In fact it is not difficult to see them as having made a very elementary mistake. Surely it is quite trivially true that whenever there are more instances of a certain property, there is at least one legitimate sense of 'more' in which there is more of that property. For example, a world in which there are more instances of roundness is _ipso facto_ a world in which there is _more roundness_. Now it is true that there is another sense in which there might be less roundness — for example, if the instances of roundness in this world are less perfect. But that does not affect the existence of one sense in which there is more. There being more instances of a certain property is, by definition, one of the ways in which there can be more of that property. If this is so, then any property that can be instantiated in different people is interpersonally additive in a limited sense. The more people who have this property, the more of that property (in one sense of 'more') there is. If more people suffer, then in one sense of 'more', there must necessarily be more suffering.[7] And if more people have the property of 'being benefited' there must, in one sense at least, be more of that property i.e. more 'benefit' or, to put it another way, more 'good'.

The point just made is obscured by Brook and Schwimmer through their choice of examples. For the examples they use are not simply properties, but degrees of properties. Thus 'height' means 'degree of vertical extension above the ground' while 'speed' means 'degree to which distance is traversed in a given amount of time'. Now when attached to nouns of this kind, the word 'more' never indicates multiplication of instances, but rather a greater degree of whatever property is involved. 'More height' means a greater degree of vertical extension above the ground, while 'more speed' means a greater distance traversed in a given amount of time. But in using these sorts of examples, Brook and Schwimmer are really changing the subject. For the sorts of abstract nouns used in the ethical discussion — e.g. 'good', 'happiness', 'suffering' — represent properties in the straightforward sense, not ones which already involve the concept of 'degree' or 'extent'. When attached to such nouns, 'more' *can* mean simply 'more instances of'. Thus it is trivially true that more instances of a property of this kind always constitutes more of that property.

But precisely *because* this is such a trivial truth, Brook and Schwimmer might reply that it is unimportant. Even if it is true that one situation's containing more instances of suffering (say) than some other entails that in some sense it has more suffering, we must provide some justification for thinking that this is a sense of 'more suffering' which provides a good *prima facie* reason for preferring the second situation to the first. Now of course it is not being claimed here that a negative utilitarian must have an *overall* preference for the second situation in such a case. He also has to take account of intensities. Later in the thesis, I want to try to explicate a *different*, stronger, sense of 'more suffering' which *does* automatically justify (from the standpoint of negative utilitarianism) an overall preference. But my concern here, for the purposes of the present argument, is only with the weaker sense of the phrase.

But in order to proceed further, we need to deepen the discussion somewhat. It was stated earlier that any property that could be instantiated in different people was interpersonally additive, in the limited sense that more people with this property will always amount to more of the property; so that, for example, if more people suffer, there must be more suffering. But this needs to be qualified. What we _should_ say is that when more people suffer _at any given moment_, there is more suffering at that moment. Without some such qualification, implausible results will emerge. We would have to say, for example, that there must be more suffering when each of five people suffer for ten seconds than when each of four suffer for two hours. But — if we assume equal intensities throughout — it does not seem that there would be _any_ sense in which there was more suffering in the first situation than in the second. The point is, of course, that suffering is instantiated in people _at times_. It might be suggested that we could construe the 'number of instances' of suffering as the number of 'person-times' at which suffering occurs. But there is great difficulty in this. 'Times' must be construed either as durationless instants or as extended intervals. If the former course is taken, there will be an infinite number of them in _any_ temporally extended situation. If the latter, we have the additional factor of the _duration_ of the intervals to worry about. Talk of discrete 'numbers of instances' of suffering does not work at all well *in general* when applied to temporally extended situations, precisely because temporal duration is itself a factor in the 'amount of suffering'. However, it can be made to work in certain special cases, and, as it happens, some of these cases provide a basis for completing the argument against Brook and Schwimmer's position.

Suppose a person with ordinary motivations has to choose between situations *A* and *B*, such that in *A* he would experience four separate one-minute periods of suffering and in *B*, five, and neither *A* nor *B* contain any other significant features. We can describe the difference between *A* and *B* as lying in the frequency with which the property of *suffering for a separate one-minute period* — or, for short, *one-minute suffering* — is instantiated. Now it is quite clear that the fact that there are fewer instances of one-minute suffering in *A* — and thus *less* one-minute suffering — gives the chooser a good *prima facie* reason for preferring *A* to *B*. And if the intensities are the same, he *will* prefer *A*. Now suppose that *A*' and *B*' are exactly like *A* and *B*, except that each one-minute period occurs for a different person (so that there are nine people involved altogether). And imagine that a utilitarian has to choose between *A*' and *B*'. The only difference between the utilitarian and the person with ordinary motivations is that he is concerned with minimizing suffering generally and not just for himself. This difference is entirely allowed for in the stipulated difference between the two pairs of situations. So if the existence of *fewer instances* of one-minute suffering in A is important for the ordinary person, why should the existence of fewer instances in A' not be important for the utilitarian? In other words, Brook and Schwimmer's position seems to involve an unjustifiable gap between *intra*personal and *inter*personal decision-making.[8]

There is besides this another serious objection that can be raised against this position. Let situations A, A' and A'' be defined as follows: in A, there are four one-minute periods of suffering, each for a different person; A' is identical to A, except for the addition of a further minute for a fifth person; and A'' also involves five minutes, but with one further difference: all the suffering in A'' is experienced by entirely *different* people from the suffering in A; otherwise A'' is identical to A'. Now suppose action a_1 would result in A and action a_2 would result in A', and there are no other relevant differences between the consequences of the two actions. This would be a case of Brook and Schwimmer's 'Pareto Exception'. a_1 would be a better action than a_2, because it would avoid the 'extra' suffering in A'. And note that although Brook and Schwimmer might not want to say that there was really 'more suffering' in A' than in A, they could hardly deny that the wrongness of a_2 was of a *consequentialist* kind. That is to say, a_2 is a worse action than a_1, because it would produce a *worse* _outcome_. But now suppose instead that while $\underline{a}_1$ would still result in _A_, $\underline{a}_2$ would result, not in _A_', but in _A_''. Brook and Schwimmer will now say that $\underline{a}_2$ is not necessarily a morally wrong action. Certainly it is not wrong in consequentialist terms, since there is (according to them) no more suffering in _A_'' than in _A_ and, unlike in the previous case, it does not come under the heading of the 'Pareto Exception'. But how can this be? Previously $\underline{a}_2$ was wrong in consequentialist terms. Now it is not wrong in those terms. But if we look at the consequences that $\underline{a}_2$ was previously imagined to have, and those that it is imagined to have now — viz. _A_' and _A_'' respectively — we find that there is _no_ relevant difference between them.[9] The only way in which these two situations differ is in the _identity_ of at least some of the suffering people — and this cannot in itself be relevant. Thus although it seems acceptable at first sight, I do not think that Brook and Schwimmer's admission of the Pareto Exception can really be made consistent with their basic position.

3. Average Utilitarianism

We saw in the opening section that the N.I.C.C. creates a challenge for someone who wishes to find a basis for the utilitarian evaluation of situations. One way of trying to meet this challenge is to take the extreme course of denying that numbers count at all. We have seen that this involves grave difficulties. Another possible route lies in Average Utilitarianism. Negative Average Utilitarianism requires us to average over the negative utility-levels of all people in each of the situations being compared, and select the situation with the lowest average.[10] If we apply it to the N.I.C.C., it will give the correct answer, since the average level of suffering in _B_ is 20, while in _A_, it is only 1 — thus making _A_ superior.

It is sometimes said that numbers have no weight at all for the Average Utilitarian. But this is an over-simplification. Numbers are an essential feature in the calculation, and thus must have an effect on the outcome. The difference between Average Utilitarianism and the extreme view considered in the previous section, according to which numbers do not count at all, can be made clear by distinguishing between the following two propositions:

(P1) If situations _X_ and _Y_ contain the same number of people, and if more suffer in _X_ than in _Y_, then, all other things being equal (i.e. same intensities and durations of suffering), _X_ must be negatively utilitarianly worse than _Y_.

(P2) If situations _X_ and _Y_ both consist of a number of suffering people, but more in _X_ than in _Y_, then, all other things being equal, _X_ must be negatively utilitarianly worse than _Y_.

The extreme view denies both P1 and P2. But Average

Utilitarianism only denies P2. For the total number of people suffering in a situation does not *as such* affect its average level of suffering. But the *proportion* of them that are suffering clearly does.

The fact that Average Utilitarianism does not deny P1 makes it less odd than the extreme view. But the fact that it does deny P2 is enough to provoke scepticism. Indeed, it seems to run directly counter to the arguments of the previous section. Thus, for example, the *X* and *Y* of P2 must be such that, for some *d*, there are more instances of suffering-for-duration-*d* in *X* than in *Y*. This means — given equality of intensity — that *X* ought to be worse than *Y*.

Even Average Utilitarianism's success with the N.I.C.C. can be seen on closer examination to be less significant than it appears at first sight. For we can produce a modified version of the N.I.C.C. in which it gives the wrong answer. Suppose situation *A* consists of one thousand people each suffering at level 1 *plus* one person *not* suffering, all throughout some interval of time *T*; while *B* consists of the same one thousand,but now relieved of suffering plus the other one now at level 20,all throughout the same interval *T*. The average in *A* will be just fractionally under 1, while the average in *B* will be just fractionally over zero. Thus for the Average Utilitarian, *B* comes out slightly better than *A*. The superiority is not as great as in the case of Total Utilitarianism, but that it obtains at all is a problem, since it is in fact much more reasonable to say (just as in the original version of the example) that _A_ is superior.

Does the Average Utilitarian have anything to say in response to these objections? In the previous chapter, we considered the so-called 'Equiprobability Model' offered by Harsanyi for the justification of Average Utilitarianism. It was criticized in that chapter as a putative justification of a _preference-oriented_ form of utilitarianism. But it could equally well be considered in a hedonistic version. Thus one might argue that the best situation, from the point of view of the negative hedonistic utilitarian, was the one that a rational self-interested negative hedonist would choose to be thrown into _at random_. Such a person would want to maximize his expected utility — construed in negative hedonistic terms — and this would require him to choose the situation with the lowest average level of suffering. If he were comparing two situations in which the durations and intensities of suffering were the same for all people involved, he would be indifferent between them, even if one of the situations contained more people. For he would be bound to end up in the same state whichever situation was chosen, given that he has to end up as _someone_ in that situation.

The oddity of the Equiprobability Model can be brought out by comparing it with an alternative model which has been suggested — the 'Superlife' Model. On the latter (in its negative hedonistic version), the correct moral choice is the one that would be made by a rational self-interested negative hedonist who supposed that he had to live in turn through _all_ the experiences of _all_ the people in whatever situation he chose.[11] On this approach, if _X_ and _Y_ are as described in P2 (page 68 above), then _X_ is worse than _Y_, since it would give a longer 'superlife' of suffering to the individual making the choice. The advocate of the Superlife Model thus affirms both P1 _and_ P2, unlike the Average Utilitarian who affirms only P1. The question is: given the option of the superlife approach, why would anyone want to adopt the Equiprobability Model? The hypothetical assumption of each model is designed to force on the self-interested person a concern for something which he would not normally be bothered about viz. the totality of _everyone's_ suffering in the situations being considered. By 'totality', I do not mean anything metaphysical, but simply the _fact_ that every one of these people is (or would be) suffering. The obvious hypothetical assumption for forcing this concern on the self-interested chooser would be the assumption that he himself would experience _all_ this suffering. Why would anyone prefer the assumption involved in the Equiprobability Model?

It seems to me that the answer to this question lies in a particular metaphysical view to which the proponent of Average Utilitarianism is implicitly giving his allegiance. This view might be termed 'relativistic solipsism'. Standard solipsism says that I alone exist. Relativistic solipsism says that just one person exists, but refuses to say whether that person is me or anyone else. Each person's complete conscious life is an alternative complete experiential reality in competition with all others to be considered as _the_ experiential reality. Thus a situation which, at the common sense level, involves a number of people each having certain experiences, fundamentally represents a number of alternative experiential viewpoints each having a claim to be regarded as _the_ true viewpoint. Each of the alternative viewpoints has an equal claim to represent the level of suffering in the situation 'as a whole'. Hence an average is used to compromise between them, as it were. The relativistic solipsist will rebut my claim that we should be concerned about the fact that _every_ person is suffering — and thus that the Superlife Model is superior — by denying that there is, literally speaking, any such fact. There is really only a range of _possibilities_ all in competition with one another to represent the suffering of a single 'indefinite' person. Although the proponent of this view may concede that more instances of suffering entails more suffering (with the qualifications discussed earlier), he will deny that if _X_ and _Y_ are as defined in P2, _X_ will be worse than _Y_. For in fundamental metaphysical terms, _X_ does not represent more instances of _actual_ suffering, but rather a wider range of _possibilities_. Since none of these possibilities creates a relevant difference between _X_ and _Y_, there will be nothing to choose between these two situations. But this is not so in the case of the _X_ and _Y_ of P1. Here a greater _proportion_ of people are suffering in _X_ than in _Y_. An adequate single 'compromise' figure for the level of suffering in _X_ will thus be higher than the corresponding figure for _Y_.

It might be thought unfair of me to foist this outlandish-seeming metaphysical view on the Average Utilitarian. But in the first place, I do not see how he can justify his preference for the Equiprobability Model over the Superlife Model except by adopting some such theory as this. And secondly, I do not think that it is really as outlandish as it appears to be. On the contrary, it seems to have a certain inherent plausibility. What it says is that experiential reality is not a totality which consists of _my_ experiences _plus_ yours _plus_ those of any other particular individual. There is no such totality. Rather, it is my experiences _or_ yours _or_ those of any other person. On this view, to talk about the fact of everyone's suffering, though useful on a common sense level, is metaphysically misleading. It would be _something_ like combining two facts from different alternative conceptual schemes and calling them a single fact. The relativistic solipsist might even argue that in trying to talk about this supposed fact, his opponent is really expressing belief in an actual superlife, instead of merely using this concept as a heuristic device. He appears to actually believe that there is one person who will live through all the experiences in question.

Although, as I have said, the relativistic solipsist's view has some plausibility, I do not believe that it can ultimately be sustained. The problem with it is that it seems to require a deep metaphysical boundary between the lives of different persons. Recent researches into the concept of personal identity, especially the work of Parfit,[12] have cast considerable doubt on the existence of such boundaries. It seems to me that the only real metaphysical or ontological boundaries within immediate experience are those between different streams of consciousness. A stream of consciousness can be defined as a maximal chain of experiences each member of which is adjacent to its successor in personal subjective time. There is thus a clear ontological separation between any two experiences belonging to different streams of consciousness — they cannot be connected by a chain of subjectively adjacent experiences. But setting the boundaries between streams of consciousness is not the same as setting the boundaries between persons. For each person's life consists of many different streams of consciousness separated (in physical intersubjective time) by periods of unconsciousness (e.g. in sleep). This suggests that if one were to adopt a relativistic solipsism at all, it would be better to frame it in terms of streams of consciousness *as opposed* to persons, for the former, unlike the latter, represent 'natural' ontological units. As far as utilitarianism is concerned, this would of course lead one to average over the utility-levels of streams of consciousness, instead of the utility-levels of persons. But relativistic solipsism seems to be much less intuitively appealing in this, than in its original, version. We have little temptation to say that each stream of consciousness is a separate alternative reality. There really *is* something called a 'life' which all these streams compose. And anyone would agree that adding to such a life an extra stream of consciousness involving suffering would necessarily make it *worse* from the standpoint of suffering. (Someone who wanted to average over streams of consciousness would, in contrast, think that it could make it *better* — if there was less suffering in the new stream than in the original ones.)

To summarize the position then: no two conscious lives are ever fundamentally more separate than two streams of consciousness. There is thus no more reason to think of different lives as constituting alternative 'worlds' than there is to think this of different streams of consciousness. But when the separate worlds are taken to be streams of consciousness, the relativistic solipsist's view becomes very unintuitive. The least uncomfortable course would therefore seem to be to abandon this view altogether. We should say instead that there is one world which contains all the experiences to which common sense attributes reality.

The relativistic solipsist's charge that in talking of the fact of everyone's suffering, we are supposing an actual superlife to exist is incorrect. Although we can think of all the individual experiential lives as jointly constituting an aggregate, more would be required than just the existence of this aggregate for one to be able to say that someone had lived the corresponding superlife. Precisely what more cannot be settled definitively because of the 'fuzzy' character of the concept of personal identity. But an obvious way of fulfilling the condition would be for the different elements of the aggregate to be correlatable with the successive states of a single human body which does not undergo any of the peculiar transformations imagined in discussions of the concept of personal identity. However, what matters from the point of view of the utilitarian moral agent is simply the intrinsic properties of the experiences in the aggregate, not any such external correlation. This is why he might just as well imagine that there was a real superlife.

4. 'Superlives' Versus 'Equiprobability': Further Discussion

M. McDermott has recently argued that, contrary to the position adopted here, the Equiprobability Model and its associated theory of Average Utilitarianism give better intuitive results than the Superlife Model.[13] McDermott presents us with a case involving a choice between situations _A_ and _B_ defined as follows: in _A_, ten people are each suffering a headache one day per week; while in _B_ one of these people is suffering a headache seven days per week and the other nine are entirely free of headaches. Many people, McDermott says, would consider such a case as a counter-example to Average Utilitarianism, since intuitively situation _A_ seems preferable to _B_, and yet there is on average less headache in _B_ than in _A_. But, McDermott argues, this fails to take into account the fact that headaches have what he calls 'increasing marginal disutility'. That is, when the frequency of headaches is already very high, a given increase in their frequency will produce a lot more extra disutility (in the way of depression and a general lowering of the sufferer's quality of life) than the same increase at a lower level of frequency. If we do take this into account, then we can see that the _overall_ harmful effect on the single headache-sufferer in _B_ of experiencing a headache seven days per week might very well outweigh the benefit to the other nine in not having any headaches at all, resulting in a higher average level of suffering for _B_ than for _A_, and so reconciling Average Utilitarianism with our intuitions. If, McDermott maintains, we ask the crucial question — viz. which of the two situations, _A_ or _B_, would we rather be thrown into at random? — our answer will probably, for this reason, be *A*.

Now suppose instead that we imagine that we have to choose between a *superlife A*, which consists of all the ten lives in the original situation *A* lived in succession, and a superlife *B*, which is similarly constructed from all the lives in situation *B*. McDermott argues that we would be most likely to choose superlife *B* over superlife *A*, since the former contains less headache on average than the latter (0.7 days per week as opposed to 1 day per week). But this shows the superlife view to be defective, since in the original example, *A* was seen to be morally preferable to *B*. Hence, concludes McDermott, the Equiprobability Model is to be preferred to the Superlife Model.

But this is very unconvincing. For exactly the same considerations which McDermott used to persuade us that Average Utilitarianism would probably view situation *A* as superior to situation *B* could also be used to argue that a rational self-interested person would probably prefer *superlife A* to *superlife B*. Why would not the extra disutility of suffering such an extremely high frequency of headaches in *one* of the constituent lives of superlife *B* outweigh the fact that the other nine lives are entirely free of headaches, thus rendering superlife *B* worse on balance than superlife *A*? Surely if situation *A* really *is* better than situation *B* in the way that McDermott suggests, then its *superlife* ought to be preferable to a rational self-interested hedonist to the superlife for B. So I fail to see that the Superlife Model is damaged by consideration of this case. Indeed the Superlife Model copes very successfully with other cases as well — such as, for example, the N.I.C.C.. Almost anyone would prefer to live through all the minor discomforts involved in A in preference to the extreme agony involved in B. Thus the superlife approach will correctly classify A as superior. In contrast, the Equiprobability Model will give the wrong answer, at least in the second version of the N.I.C.C. described on page 69, where situation B is expanded so as to include a vast number of people who are not suffering. In such a case, a rational self-interested hedonist might well prefer to be thrown at random into B rather than A, since he might figure that the probability of his ending up as the person experiencing agony would be too small to be worth worrying about.

With the adoption of the Superlife Model we carry out a complete assimilation of interpersonal decisions to intrapersonal ones. Persons are no longer significant units — at least as far as the utilitarian is concerned.[14] To underline this point, I want now to replace the N.I.C.C. with a different case, which I shall refer to as the 'T.I.C.C.' (for 'Time-Intensity Conflict Case'). In this case, situation A simply involves one thousand minutes of suffering at level 1 and situation B involves just one minute of suffering at level 20. The T.I.C.C. does not stipulate any particular distribution of the different minutes of suffering between different people, since this is not considered a significant factor for utilitarian calculation. But the question still remains with respect to the T.I.C.C., as with the original N.I.C.C.: can a utilitarian account for our feeling that _A_, and not _B_, is superior, and if so how?

5. The Abandonment of Absolute Disvalues

We have decided that the best approach for making a utilitarian comparison between situations is the Superlife Model. The adoption of the Superlife Model renders certain possible solutions to the quantification problem — particularly Average Utilitarianism — highly implausible. But it clearly does not tell us immediately what the correct solution _is_. The Superlife Model tends to be associated with Total Utilitarianism. Clearly it does fit better with that theory than with Average Utilitarianism. But the fatal problem with Total Utilitarianism is that it gives the wrong result in the N.I.C.C. and T.I.C.C.. The superlives which it would require us to choose are not those which, _qua_ rational hedonists, we should choose.

I believe that it would be a mistake to look for a solution to the problem of quantitative comparison _on the same general lines_ as Average Utilitarianism and Total

Utilitarianism. By this I mean that we should not suppose that each situation has an absolute value or disvalue, this being determined by a *function* which takes as its arguments the values of the individual utilities and disutilities in the situation, and such that the best situation is determined by comparing the values of the function for each alternative. I shall offer two arguments against the possibility of assigning absolute disvalues to situations consistently with our intuitions:

(1) Suppose $\underline{A}_1$ involves one minute of suffering at level 5, while $\underline{A}_2$ involves one minute of suffering at level 10. How will the absolute disvalue of $\underline{A}_2$ compare with that of $\underline{A}_1$? It seems difficult to avoid saying that it is double. What else *could* one say? Now suppose that $\underline{A}_3$ involves two minutes of suffering at level 5. How does the absolute disvalue of $\underline{A}_3$ compare with that of $\underline{A}_1$? Again any answer other than saying that it is twice as great seems unreasonable. And in general, it appears that for any *n*, multiplying either duration *or* intensity by *n* will necessarily mean multiplying the absolute disvalue by *n*. Any increase in duration by a given factor is precisely as bad as an increase in intensity by the *same* factor. But this seems to be tantamount to Total Utilitarianism in which we simply sum over equal periods of constant intensity. Certainly it seems to contradict the intuitions which underlie our reaction to the T.I.C.C.. If we multiply the intensity of a minor discomfort by 20, we

will get something *much worse* than if we multiply its duration by 20. In other words, once we take the step of talking about absolute disvalues in the *first* place, we seem bound to end up contradicting this intuition.

(2) Even if we do not feel that we have to say that multiplying the duration of a one-minute minor discomfort by *n* necessarily means multiplying its absolute disvalue by *n*, we still have to give an account of how increasing its duration *does* affect its disvalue. And it is very difficult to do this consistently with our intuitions. For it seems in fact that *no* increase in duration can be as bad as a very large increase in intensity. And yet at the same time, increasing the duration must presumably always lead to *some* raising of the disvalue. The only way to reconcile these two facts seems to be to say that the duration of minor discomforts is affected by diminishing marginal disutility, so that a given increase in duration is much less significant when the duration is already very high than when it is lower. But this would imply that eventually the extra disvalue of an additional minute of discomfort would be so small as to have no practical significance. But if minor discomfort is worth worrying about at all, then surely it must always be worth avoiding an *extra* minute of it (when there are no additional costs to consider).

These two arguments should convince the reader that we need an alternative to the absolute disvalue approach.

Just what this alternative could be is the subject of the next chapter.

ENDNOTES

[1] The theoretical difficulties involved in doing this in an exact way are discussed in Chapter Four, pp. 92-93, and in Appendix A, endnote 5, pp. 157-158. There is no doubt that the operation can be carried out to any desired degree of accuracy.

[2] Writers who have recently discussed this kind of example, and come to the same conclusion about it, include Richard Brook and Seymour Schwimmer, "On Adding the Good", Social Theory and Practice, 7 (Fall, 1981), p. 328 (this article will be considered further in Section Two) and Marilyn McCord Adams, "Hell and the God of Justice", Religious Studies, 11 (1975), p. 439. The belief that situation *B* is superior bears a close relation also to Parfit's 'repugnant conclusion'. (See Derek Parfit, Reasons and Persons (Oxford: Clarendon Press, 1984), p. 388.)

[3] Strictly speaking, the application of these arguments requires a slight change of perspective. One needs to consider the question, not simply of which of the two situations is intrinsically better, but which of two possible alternative actions *a* and *b*, that would lead to *A* and *B* respectively, ought to be done. One needs to suppose in fact that the suffering in *A* and *B* represents all and only the suffering whose existence or non-existence is causally dependent on which of *a* or *b* is done, and that *a* and *b* are also indifferent with respect to other possible consequences besides suffering. One can then argue that doing *b* would be unjust, since it would inflict extreme suffering on someone by an action whose only merit would be to spare a much larger number of people a very trivial harm; or that doing *b* would lead to unacceptable inequality — the inequality between the person at level 20 and those who, through the choice of *b*, are spared all suffering during the time at which *A* would have obtained had *a* been done instead.

[4] I would like to mention here a recent argument which claims to prove Total Utilitarianism from very weak assumptions. It appears in Yew-Kwang Ng and Peter Singer, "An Argument for Utilitarianism", Canadian Journal of Philosophy, 11 (June, 1981), pp. 229-239. Their crucial premise, which they call the principle of 'Weak Majority Preference' states

that 'for any community of _n_ individuals choosing between two possibilities _x_ and _y_, if no individual prefers _y_ to _x_ and at least _n_/2 individuals prefer _x_ to _y_, then _x_ increases social welfare and is preferable'. (_Ibid_., p. 232.) Briefly, the authors try to use this principle to show that for any two possible states of society $\underline{S}_1$ and $\underline{S}_2$ which are equal in total utility, one can construct a series of states, beginning with $\underline{S}_1$ and ending with $\underline{S}_2$, such that each member of the series can be seen to be equal in social welfare to its predecessor, thus proving (by transitivity) that $\underline{S}_1$ is equal in social welfare to $\underline{S}_2$. Thus Ng and Singer would argue, for example, that if $\underline{S}_2$ is derived from $\underline{S}_1$ by lowering the welfare of one individual by 20 units and raising the welfare of each of twenty people by 1 unit, one must be able to construct an appropriate series to show that $\underline{S}_1$ would be equal in social welfare to $\underline{S}_2$, contrary to the position adopted here. The crucial feature of these series which enables Ng and Singer to argue that they leave social welfare unchanged, is that in going from each member to its successor, all the 'harms' inflicted are so small as to be below the threshold of perception of those affected. This means that these people could not actually be said to _prefer_ the previous situation to its successor and thus, by Weak Majority Preference, or at least a closely related principle, the successor is no worse a situation. But this is highly questionable. If I undergo a deterioration in welfare which I _could_ have noticed, but contingently did not do so, this fact would not mean that the situation had not, to that extent, been worsened. So why should it not have been worsened when my failure to notice is (physically) necessary instead of merely contingent? Of course, if the change is so small, we cannot say that the worsening is practically significant, but that there has _been_ a _prima facie_ worsening seems undeniable. This means that — theoretically — to know the 'net' effect, one must balance this _prima facie_ worsening (and any others) against any _prima facie_ improvements. But that was exactly what Ng and Singer (correctly) wished to avoid. The balancing-operations of Total Utilitarianism are precisely what provoke scepticism.

[5] Brook and Schwimmer, "On Adding the Good", _Social Theory and Practice_, 7 (Fall, 1981), pp. 325-335. A somewhat similar position has also been argued for by John Taurek in "Should the Numbers Count?", _Philosophy and Public Affairs_, 6 (Summer, 1977), pp. 293-316. Taurek's article is discussed by Derek Parfit in "Innumerate Ethics", _Philosophy and Public Affairs_, 7 (Summer, 1978), pp. 285-301.

[6] Brook and Schwimmer, "On adding the Good", p. 327.

[7] This statement is qualified later. See below, p. 64.

[8] This argument is also advanced by Gregory Kavka in his critique of the Taurek article mentioned in note 5. (Gregory Kavka, "The Numbers Should Count", Philosophical Studies, 36 (October, 1979), pp. 285-294 — see especially pp. 292-293.)

[9] By a principle adopted in the next chapter — 'NCP1' — A is negatively utilitarianly superior to A' and to A'' for precisely the same reason in each case: that one can map each period of suffering in A onto a (different) period of suffering of equal intensity and duration in A' or A'' and still have suffering 'left over' in these situations. (See pp. 91-92.) The fact that, in the case of A and A'', experiences of different people are mapped onto one another is unimportant.

[10] When calculating the average, all people involved in the situation must of course be included, whether they are suffering or not. They can be assigned an intensity of zero, if they are not in fact suffering. This is unnecessary for the application of Total Utilitarianism, and also for the theory proposed in Chapter Four.

The theory I am describing here averages over persons. But how is the utility-level of an individual person decided? One could average over all the moments of that person involved in the situation, but it seems that one also has the option of summing over these moments as Total Utilitarianism does. Also, there is the possibility of an Average Utilitarianism which does not average over persons, but sums over persons and averages over times. These and other distinctions between different kinds of Average Utilitarianism are investigated in detail in T. M. Hurka, "Average Utilitarianisms", Analysis, 42 (March, 1982), pp. 65-69, and "More Average Utilitarianisms", Analysis, 42 (June, 1982), pp. 115-119. Hurka shows that every variety which he considers is counter-intuitive in some respect.

[11] The two models are contrasted in M. McDermott, "Utility and Distribution", Mind, 91 (October, 1982), p. 577. (This article will be discussed further below.) McDermott mentions R. M. Hare as one of the proponents of the Superlife Model. He refers to Hare's Freedom and Reason (Oxford: Clarendon Press, 1963), p. 123. As its name suggests, the Superlife Model involves imagining that one is living through a succession of what are really in themselves complete lives. I have adapted this model to the peculiarities of my own discussion, by supposing that a 'superlife' corresponds to a specific situation which may not contain the experiences of an entire life.

[12] See Parfit, Reasons and Persons, Part Three, pp. 199-347.

[13] McDermott, "Utility and Distribution", p. 577.

[14] Of course, this has often been considered an _objection_ to utilitarianism. Thus John Rawls accuses the utilitarian of 'conflating all persons into one through the imaginative acts of the impartial sympathetic spectator' and observes that 'utilitarianism does not take seriously the distinction between persons' (John Rawls, _A Theory of Justice_ (Cambridge, Massachusetts: Harvard University Press, 1971), p. 27). But I believe that the arguments of this section show why this is justified. It is interesting to note, incidentally, that Sidgwick argued for utilitarianism partly by drawing an analogy between interpersonal _morality_ and intrapersonal _prudence_. See Henry Sidgwick, _The Methods of Ethics_ (7th. ed.; London: MacMillan and Co., 1907), p. 418.

CHAPTER FOUR

CORRELATION AS THE BASIS OF UTILITARIAN SUPERIORITY

In this chapter, I switch from the negative, critical stance adopted in Chapter Three to the positive task of finding an adequate theory of the conditions under which one situation may be said to be utilitarianly superior to another, whether positively or negatively. I shall continue to talk about negative superiority, although the theory that I shall produce has a structurally identical version dealing with positive superiority. The question then is this: under what conditions may a situation be said to be superior to another _from the standpoint of suffering_? — i.e. when may the first be said to 'involve less suffering' than the second?

1. Synopsis of the Chapter

The unusual aspect of the method proposed in this chapter is that it involves not a single principle of choice, but rather a plurality of principles of varying degrees of validity. I call the principles 'correlative' because they involve correlating experiences in one situation with experiences in another. An infinite hierarchy of what I call _perfect correlative principles_ is introduced, each member of which is more complex than its predecessor. However, since the logic of each principle is simply an extension of that involved in its predecessor, it is natural to suppose that all the members of the series have _equal_ validity, and thus that they can be combined into a _single_ general principle of (negative) correlative superiority. A formulation of this general principle is suggested. But then in the following section a serious difficulty emerges. It is shown that the general principle gives the _wrong_ result in the T.I.C.C.. It judges superior situation _B_ in which there is one minute of agony, instead of situation _A_ in which there are one thousand minutes of minor discomfort. In fact, it turns out that the general principle collapses into the Principle of Total Utility, a principle already rejected as highly implausible. How do we escape from this _impasse_? The only possibility, it appears, is to abandon the general principle and return to the original _series_ of principles from which it was derived, _denying_ that each member of the series has equal validity, but supposing instead that each has a degree of validity more-or-less proportionate to its intuitive plausibility. The first two principles — which I call 'NCP0' and 'NCP1' — have the greatest validity. The next — NCP2 — is not as valid as these two, but it is still very acceptable. By the time we get to NCP3, however, the _complexity_ involved is such as to make the matter somewhat doubtful, and this doubt increases as we ascend the hierarchy, until we get to principles that have virtually no force at all. And fortunately, in order to prove that _B_ is superior to _A_ in the T.I.C.C., we would need NCP20, which is far too complex to have any real validity.

In the last section of the chapter I introduce a new hierarchy of principles called _imperfect correlative principles_. Each perfect correlative principle — i.e. each member of the original hierarchy already referred to — has an imperfect principle corresponding to it. Each imperfect principle has more validity than its successor in the series, but slightly _less_ validity than its corresponding perfect principle. However, a very low-order imperfect principle takes priority over a very high-order perfect one. It is the imperfect principle INCP0 — corresponding to the perfect principle NCP0 — which produces the result that _A_ is superior to _B_ in the T.I.C.C.. As we have already mentioned, according to NCP20, _B_ is superior to _A_. But because INCP0 is so much simpler, and thus more acceptable, than NCP20, the former should be adhered to rather than the latter, despite the fact that it is only an imperfect principle. The loss in validity due to 'imperfection' is outweighed by the gain in validity due to simplicity. This is the reason why _A_ is preferable to _B_ in the T.I.C.C..

2. The First Two 'Perfect Principles'

Suppose that we have two situations _X_ and _Y_, such that there is no suffering in _X_ and some suffering in _Y_. Then obviously _X_ is negatively utilitarianly superior to _Y_.

The first member of our hierarchy of perfect principles — called 'NCPO' — deals with this case only. It can be stated simply thus:

> NCPO(S)[1]
>
> For any *X* and *Y*, if *X* contains no suffering, and *Y* contains at least some suffering, then *X* is negatively utilitarianly superior to *Y*.
>
> ('*X*' and '*Y*' range over possible situations.)

Of course this principle does not actually involve any *correlation*, since there is nothing in *X* to correlate. This is why it is assigned the number zero. It represents a kind of limiting case.

Now suppose that in *X* there is one minute of suffering at intensity 5, and in *Y* two minutes of suffering also at intensity 5. Call this 'Case (a)'. It is obvious which situation should be chosen in Case (a) — situation *X*. Suppose alternatively (Case (b)) that in *X* there is one minute of suffering at level 5 and in *Y* one minute at level 6. Again it is obvious that *X* should be chosen in preference to *Y*. And the reasoning which justifies the preference for *X* is very similar in the two cases. The point is that *Y* has *at least as much suffering and more* than *X*, not in the rather abstract sense of (say) the Principle of Total Utility, but rather in the straightforward sense that the minute of suffering in *X* can be matched with a minute in *Y* which is at least as intense (showing that *Y* has *at least as much* *suffering* as *X*) and *in addition*, either there is further suffering in *Y* *beyond* this minute (as in Case (a)) or this minute is more intense than the minute in *X* (as in Case (b)) (showing that *Y* has *more* suffering than *X*).

My second 'perfect principle' — NCP1 — is intended to represent the general form of this type of reasoning. First of all, it would seem that the periods of suffering into which the two situations are divided can be of any duration. Let us use the term *segment* to refer to any period, or aggregate of non-successive periods, of any duration, within a 'superlife' formed from the suffering within the two situations. Any division of two situations into segments in this sense will be referred to as a *segmentation*. Using this terminology, we may state NCP1 thus:

> *NCP1(S) (Version A)*[2]
>
> For any *X* and *Y*, if the following condition is satisfied, then *X* is negatively utilitarianly superior to *Y*:
>
> There is a segmentation *S* of *X* and *Y*, such that all the segments under *S* are of constant intensity and the following conditions are satisfied:
>
> (A) There is a one-to-one mapping *M* of all *X*-segments under *S* onto some subset of all *Y*-segments under *S*, such that for any *X*-segment σ , *M*(σ) is equal in duration to

$\underline{\sigma}$ and at least as intense,

AND, IN ADDITION:

(B) EITHER:[3]

(B1) for at least one *X*-segment $\underline{\sigma}$, *M*($\underline{\sigma}$) is *more* intense than $\underline{\sigma}$,

OR:[3]

(B2) there are *Y*-segments under *S* not correlated by *M*.

One difficulty with the above principle is that some, if not all, of our experiences may be temporally continuous rather than discrete, which means that the maximal units of constant intensity might in some cases be durationless *instants* rather than enduring periods, and we cannot allow segments to be mere instants. For if we did, the mapping *M* could not always be relied upon to do the job intended for it. Suppose *X* contains a pain which increases continuously from intensity 0 to intensity 5 in two minutes, while *Y* contains a pain which does the same in just one minute. Since there are the same number of instants in any period of time of whatever length, we can construct a one-to-one mapping *M* which correlates each instant of pain in *X* with an instant of equally intense pain in *Y*. Now suppose we add a few seconds of *extra* pain to *Y*. It is clear that if instants can count as segments, then clauses (A) and (B2) of version A of NCP1(S) are satisfied in this case, despite the fact that there is actually a *greater* duration of suffering in *X* than in _Y_. NCP1 was obviously not intended to give this kind of result.

Some might maintain that no experience can really be temporally continuous. They might argue that variations in our experiences that are too small to be noticeable have no _phenomenological_ reality. (Thus according to such people, when I hear a clarinet note with what seems like a continuously changing pitch, I am not really hearing infinitely many different pitches, but only some finite number.) This argument strikes me as dubious. It seems to me that there _can_ be real differences in our experience that we do not notice and could not notice, owing to the limits of our introspective capacities. But the question is far too difficult and far too remote from our central concerns to detain us here. The point is that in order to be _sure_ of having given a completely general account of correlative superiority, we must allow for the _possibility_ that some of our experiences are temporally continuous.

So how should we state NCP1 for the case of temporally continuous experience? It is clear, in the first place, that the statement should refer to experiences _extended in time_, since we have no clear intuitions about instants as such. The suggestion which follows uses the very simple notion of a _descendant_ of a segmentation. One segmentation is a descendant of another when it is derivable from that other merely by dividing at least one of its segments into segments of shorter duration.

> NCP1(S) (Version B)
>
> For any *X* and *Y*, if the following condition is satisfied, then *X* is negatively utilitarianly superior to *Y*:
>
> There are segmentations $\underline{S}_1$ and $\underline{S}_2$ of *X* and *Y* respectively whose combination satisfies Version (A) of NCP1(S) (with 'intensity' read as 'mean intensity'), and for any descendant $\underline{S}_1$' of $\underline{S}_1$, there is some descendant $\underline{S}_2$' of $\underline{S}_2$, such that the combination of $\underline{S}_1$' and $\underline{S}_2$' does so too.

More informally, if we can indefinitely continue the process of producing finer and finer segmentations all of which satisfy the relevant condition, then the kind of superiority we are concerned about here does indeed obtain. Talk about 'instants' is really just elliptical for talk about the theoretical possibility of such an infinite process. It should be noted, however, that for all *practical* purposes, the first version of the principle is perfectly adequate, even for temporally continuous experience. All we need do to apply it in the latter case is ensure that our segments are small enough to have *approximately* constant intensities. It is for this reason that I shall henceforth refer only to the *A*-version of each principle discussed (i.e. the one presupposing discreteness).

Up to now we have spoken only of the *superiority* of one situation over another. But there is also another concept of very great importance viz. that of one situation's being *equal* in value or disvalue to another. In the present case of negative utilitarian comparisons there must be conditions for situation *X* to have *exactly as much suffering* as situation *Y*. One such condition has an obvious logical connection with the superiority-condition NCP1(S) already stated:

> *NCP1(E)*
>
> For any *X* and *Y*, if the following condition obtains, then *X* is negatively utilitarianly equal to *Y*:
>
> There is a segmentation *S* of *X* and *Y*, such that all segments under *S* are of constant intensity, and there is a one-to-one mapping *M* of all *X*-segments under *S* onto all *Y*-segments under *S*, such that for each *X*-segment σ , *M*(σ) is of equal duration to σ and of the same intensity.

The full NCP1 incorporates both NCP1(S) *and* NCP1(E).

There is an important, and intuitively obvious, relationship between superiority and equality which holds not just at the NCP1-level, but at *all* levels of the hierarchy. It can be expressed in the following rule: if *X* is equal in suffering to *Y* and *Y*' is obtained from *Y* simply by increasing the intensity of some of the suffering in *Y*, or adding extra periods of suffering to *Y*, or both, then *Y*'has more suffering than *X*. Thus whenever *X* is equal to *Y* by the E-version of any perfect correlative principle, it follows that *X* is *superior* to *Y*' by the S-version of the same principle.

Let us now return to NCP1. It is important to notice how strong it is. It is stronger, for example, than the Pareto Principle, according to which a change which benefits at least one person and harms no-one represents an improvement in overall welfare. If we added to NCP1(S) the stipulation that the mapping *M* only pair off segments that are *experienced by the same person*, we would get a weaker principle that was entailed by a negative hedonistic version of the Pareto Principle. Suppose *X* has one segment of intensity 5 and duration one minute and *Y* has two such segments. If one of the *Y*-segments is experienced by the *same* person as the single *X*-segment, then by this weaker principle, *X* is undoubtedly better than *Y*, since at least one person is better off, and no-one is worse off. If, on the other hand, neither *Y*-segment belongs to the same person as the *X*-segment, then, whether we choose *X* or *Y*, *both* situations are such that someone is worse off in them than he is in the other, and so the weaker principle does not favour either situation. But NCP1(S) does, since it does not impose this identity-restriction on *M*. Thus according to NCP1(S), one situation may be better overall than another, even though some people are worse off in the first than they are in the second. But this does not seem to be a defect in NCP1(S). On the contrary, when the conditions laid down in NCP1(S) are satisfied, it seems intuitively clear that _X_ _is_, all things being equal, superior to _Y_, and thus, at least when there are no competing reasons for preferring _Y_, _X_ should be preferred.

The application of NCP1 can be illustrated by reference to the kinds of examples discussed by Parfit, in which our decisions affect the identity of future generations.[4] Suppose policy _p_ will result in the existence, during some period of time _T_, of some population all experiencing a certain very low quality of life, while policy _q_ would result in the existence, during the same period, of a smaller population of totally _different_ people all enjoying a much higher quality of life.[5] Then Parfit would maintain — and many would agree — that, as far as its effects during _T_ were concerned, _q_ was a much better policy than _p_. And this is despite the fact that the people who will exist during _T_ if _q_ is carried out could not be said to benefit from the choice of _q_, since if _p_ had been carried out instead, none of them would even have existed. In order for NCP1 to be applicable to this example, we need to understand 'quality of life' in negative hedonistic terms. Then it is highly likely that a superlife constructed from the suffering of the people who would exist during _T_ if _p_ is followed and a superlife constructed from the suffering of the people who would exist during _T_ if _q_ is followed would be such that one could construct a segmentation and a mapping which, according to the conditions laid down in NCP1(S), would show the latter superlife to be better than the former. One would be able to find, for each segment in the latter, a segment in the former with at least as great an intensity of suffering, and in many cases greater. Thus from the point of view of their effects during _T_, NCP1 would judge _q_ to be a better policy than _p_.[6] And this is despite the fact that the choice of _q_ cannot be said to (negatively) benefit those who will exist during _T_ if _q_ is followed. Indeed, it negatively _harms_ them, since if _p_ had been followed instead, they would not have existed and so would not have suffered at all. Thus minimizing suffering is not necessarily a matter of causing people to suffer less than _they_ (those very people) otherwise would have. As in the abstract example on page 96 , we see that the correlations which it involves do not have to reflect trans-situational personal-identity-relations. This is of course entirely in accordance with my advocacy, in Chapter Three, of the 'superlife' approach in which the boundaries between different people's lives are disregarded.

3. A Third Correlative Principle

It is fairly easy to show that even if the combination of NCP0 and NCP1 is a sufficient condition for negative utilitarian superiority, it is not a necessary one. Consider the following choice-situation:

SITUATION A	SITUATION B
A one-minute pain of intensity 1	A two-minute pain of intensity 8
A one-minute pain of intensity 9	

Which of these two situations — *A* or *B* — is preferable? First of all, this case obviously does not satisfy NCP0, since there is suffering in both situations. Furthermore, it does not satisfy NCP1 either. *B* cannot be superior to *A* by NCP1, since it is not possible to match up *both* minutes of pain in *B* with a minute of at-least-equally-intense pain in *A*. And *A* cannot be superior to *B* by NCP1, since the pain of intensity 9 in *A* cannot be appropriately matched up. And yet it does seem pretty clear that *A* is superior to *B*. This is because, by matching up the pains in *A* and *B* appropriately, we can think of the choice between the two as involving, on the one hand, having a 1 instead of an 8, and, on the other, having an 8 instead of a 9. Since 8 minus 1 is much greater than 9 minus 8, the first of these alternatives seems greatly superior to the second, and thus the situation containing the 1 — situation A — should be chosen, despite the fact that it actually contains the pain of highest intensity. To make the reasoning a little more concrete, imagine that one had to choose between, on the one hand, a situation involving both a very severe headache and a period of mild discomfort of duration exactly equal to that of the headache and, on the other, a situation involving two headaches, each of the same duration as before, and only slightly less severe than the headache in the first situation. It seems reasonable to suppose that one would choose the first situation, despite the fact that the worst headache occurs in it, and not in the second.

To return to our abstract terminology, we may say that the reason why _A_ is superior to _B_ is that one can divide _A_ and _B_ into segments, such that there is a one-to-one mapping _M_ of each _A_-segment onto a _B_-segment of equal duration, and also such that for each _pair_ in _M_ in which the intensity of the _A_-member is greater than that of the _B_-member (and which therefore represents a respect in which _B_ is superior to _A_), we can find another pair in which the intensity of the _B_-member is greater than that of the _A_-member and _by a higher margin_. This shows that if we want to give a completely comprehensive account of correlative superiority, we must consider, not only mappings of segments onto other segments, but also of pairs _within_ such mappings onto other such pairs. For this purpose, given a segmentation of any two situations _X_ and _Y_, and a one-to-one mapping of _X_-segments onto _Y_-segments, we divide the pairs belonging to the mapping into two overlapping classes: _intensifications from X to Y_ (or 'X-to-Y intensifications'), in which the member from _Y_ has at least as great an intensity as the member from *X*; and *intensifications from Y to X* (or 'Y-to-X intensifications') in which the *X*-member has at least as great an intensity as the *Y*-member. The *magnitude* of an intensification is simply the difference between the intensities of its two members. Zero-magnitude intensifications represent limiting cases of both *X*-to-*Y* *and* *Y*-to-*X* intensifications. Given this terminology, we can state our third perfect correlative principle — NCP2 — as follows:

> *NCP2(S)*
>
> For any *X* and *Y*, if the following condition is satisfied, then *X* is negatively utilitarianly superior to *Y*:
>
> There is a segmentation *S* of *X* and *Y* such that all the segments under *S* are of constant intensity and the following conditions are satisfied:
>
> (A) There is a one-to-one mapping M_1 from a subset of the *X*-segments under *S* onto a subset of the *Y*-segments under *S*, such that for any *X*-segment σ , M_1(σ) is equal in duration to σ ,
>
> AND, IN ADDITION:
>
> (B) at least some of the pairs constituting M_1 are correlated by a further one-to-one mapping M_2 of which the following is true:

(B1) The domain Δ of M_2 is the set consisting of any non-zero Y-to-X intensifications in M_1 plus any X-segments not correlated by M_1.

(B2) Its range Δ' is some subset of the set β consisting of any non-zero X-to-Y intensifications in M_1 plus any Y-segments not correlated by M_1.

(B3) For any member I of Δ, all the segments involved in[7] I or $M_2(I)$ (the correlate of I under M_2) are equal in duration and $M_2(I)$ is at least as great in magnitude or intensity as I.

(B4) EITHER:

(B4.1) for some member I of Δ, $M_2(I)$ is *greater* in magnitude or intensity than I,

OR:

(B4.2) Δ' is a *proper* subset of β i.e. there are members of β not correlated by M_2.

By comparing this principle with Version A of NCP1(S), the reader will note clear similarities. However the principle is slightly more complicated than might be expected. I require that the mapping M_1 be from a *subset* of the X-segments (and not from the whole set as in the case of NCP1), and then I include in Δ , the domain of M_2, not only non-zero *Y*-to-*X* intensifications in M_1, but also *X*-segments not correlated by M_1 (i.e. X-segments not involved in any of the pairs belonging to M_1). And similarly I include in Δ', the *range* of M_2, not only non-zero *X*-to-*Y* intensifications belonging to M_1, but also *Y*-segments not correlated by M_1. To see why this is done, consider the following case:

SITUATION A	SITUATION B
One minute at intensity 5	One minute at intensity 10
One minute at intensity 3	

We can see this choice-situation as involving a 5-to-10 intensification from *A* to *B* plus a pain of intensity 3 in *A*. *A* can be seen to be superior to *B* by noting that the 'extra' pain in *A* can be mapped onto the intensification from *A* to *B*, whose magnitude is greater than the intensity of that pain. In other words, *X*-segments should count as equivalent to non-zero *Y*-to-*X* intensifications, and similarly *Y*-segments should count as equivalent to non-zero *X*-to-*Y* intensifications.[8] This, then, is the reason for the extra complexity in the specification of the domain and range of the mapping M_2.

Note that a Total Utilitarian would agree with the advocate of NCP2 that *A* was superior to *B* in the above case. For 5 + 3 is less than 10. It might even seem as if one who uses NCP2 is simply using the Principle of Total Utility in disguise. For remember that the key point in the application of NCP2 to this example was that 3 was less than 10 - 5 i.e. that 5 + 3 is less than 10! But it is not really true that this is simply an application of the more conventional principle. For although it employs the same arithmetic, the *rationale* underlying the use of this arithmetic is different. This rationale involves a much weaker condition than the Total Utility Principle. According to the latter, if we add an extra minute at intensity 1 to situation *A*, *A* is still superior to B, but we cannot prove superiority by NCP2 in that case. The Total Utility Principle is based on the idea that, for any *n*, multiplying the duration of suffering by *n* is exactly as bad as multiplying the *intensity* of suffering by *n*. The user of NCP2, however, is only committed to the truth of this statement when *n* is equal to 2. Suppose, for example, *A* and *B* are as follows:

SITUATION A	SITUATION B
One minute at intensity 20	Two minutes at intensity 10

Here there is in *A* suffering twice as intense, but lasting only half as long, as in *B*. By NCP2 (E) — i.e. the equality-principle corresponding to the *superiority*-principle (NCP2 (S)) already stated — *A* is exactly as bad as *B*. Thus there is clearly a close relationship between the Total Utility Principle and NCP2. They are nonetheless *different* principles.

4. The Hierarchy of Perfect Principles and the General Principle NCP

The reader will probably be able to imagine the rough character of NCP3 (which I shall not formally state). Just as NCP2 takes the pairs that belong to a mapping of segments onto segments, and maps *these* pairs onto one another, so NCP3 maps pairs of pairs onto pairs of pairs. However we should note that, just as in the case of NCP2, where a single segment may take the place of a pair of segments (see the example on page 103), pairs of segments and single segments may count as pairs of pairs for the purposes of NCP3. Consider the following modification of the example on page 103:

SITUATION A	SITUATION B
One minute at intensity 5	One minute at intensity 10
One minute at intensity 3	
One minute at intensity 1	

We saw that in the original example, *A* was superior to *B* by NCP2, since 10 - 5 is greater than 3. We should now be able to see that the new *A* is also superior to *B* by NCP3, since (10 - 5) - 3 is greater than 1. But the reader may have difficulty in seeing this as an application of NCP3, and further explanation is called for. First we should note that the suffering in *A* and *B* is of *unequal total duration*. This means that if we choose, we can fill in the description of *B* with references to moments during which no suffering exists. These may be represented by zeros:[9]

SITUATION A	SITUATION B
One minute at intensity 5	One minute at intensity 10
One minute at intensity 3	One minute at intensity zero
One minute at intensity 1	One minute at intensity zero

Now suppose we consider an additional minute of suffering which is the same whichever of *A* or *B* is chcsen. Suppose that throughout this minute the intensity of suffering is 5. There is nothing to stop us adding this minute to the description of *A* and *B*:

SITUATION A	SITUATION B
One minute at intensity 5	One minute at intensity 10
One minute at intensity 3	One minute at intensity zero
One minute at intensity 1	One minute at intensity zero
One minute at intensity 5	One minute at intensity 5

We end up with exactly the kind of case that NCP3 is designed for — one which involves the mapping of a pair of pairs onto another pair of pairs. The bottom two rows represent a pair of pairs which, by NCP2-type reasoning, favours *B* over *A*, since 1 minus zero is greater than 5 minus 5. The top two rows, however, represent a pair of pairs that favours *A* over B, since 10 minus 5 is greater than 3 minus zero. But since the *degree* of superiority for *A* represented by the top two rows (the difference between 10 - 5 and 3 - 0) is greater than the degree of superiority for *B* represented by the bottom two (the difference between 1 - 0 and 5 - 5), *A* comes out as superior *overall*.

If the reader is still not convinced that *A* is superior to *B* according to NCP3, he can if he wishes imagine the zeros in *B* replaced by *negligibly small pains* and the 5 in *B* replaced by pain which is negligibly *more* intense than 5. This means that we are dealing with real pains throughout and that the bottom row represents a 'real' intensification from *A* to *B* paralleling the 5-to-10 intensification in the top row. Now there seems to be no reason at all to refuse to apply NCP3. But since the changes made are only negligible, and since NCP3 can clearly be applied *after* they are made, why shouldn't it also be applicable *before*?

It should be fairly obvious to the reader that we can construct an infinity of principles on the lines of NCP0-3, each differing from the others in its precise degree of complexity. For any $n > 1$, NCP*n* *primarily* correlates pairs that involve *n* - 2 *sub-nestings* of constituent pairs before we reach the level of actual segments. (For example, NCP4 correlates pairs that involve *two* such sub-nestings.) But as we have seen, pairs of *lesser* complexity and even segments can be permitted to 'stand proxy' for such pairs.

Once having launched ourselves on the journey up the hierarchy of perfect principles, it is difficult to find anywhere to stop. Why, for example, should we accept NCP2 and yet reject NCP3? Or again, why give NCP3 our seal of approval as a valid utilitarian principle, but draw the line at NCP4? This suggests that we should consider either *none* of the principles to be valid, or else *all* of them. Since at least some of the principles seem to commend themselves to our intuitions (I am hoping the reader agrees with me on this point), it would appear that we cannot take the former course. Therefore the latter seems to be required. That is to say, it seems to be required that we affirm each member of the infinite hierarchy as a sufficient condition for the negative utilitarian superiority of a situation *X* over a situation *Y*. The principles can in fact be *combined* into a single general principle of correlative superiority, which I shall refer to as 'NCP'. Before we can state NCP, we need to generalize the notion of an 'intensification' used in the statement of NCP2. This is done in the following inductive definition:

Definition of an 'intensification from X to Y'

1. An *intensification from X to Y of order zero* is a *Y*-segment. Its *magnitude* is its intensity.

2. An *intensification from X to Y of order n* is a pair composed of a *Y*-to-*X* intensification *I* and an *X*-to-*Y* intensification *J* such that:

(a) One of *I* or *J* is of order *n* - 1 and the other is of order *n* - 1 *or less*.

(b) The magnitude of *J* is greater than or equal to the magnitude of *I*.

(c) No segment is used in the construction of both *I* and *J*.

(d) All segments used in the construction of either *I* or *J* are equal in duration.

The *magnitude* of such an intensification is the difference between the magnitudes of its two members.

(N.B. A segment used in the construction of an intensification will in future sometimes be referred to as an *ultimate constituent* of that intensification.)

Note how the *order* of an intensification, which is a measure of its complexity, is determined. Any intensification of order *n*, where *n* is greater than zero, must be a pair of simpler objects, one of which is an intensification of order exactly *n* - 1. Because of this, such an intensification must have *n* lower levels discernible *within* it. (For example, an intensification of order 1 must have *one* level — that of individual segments — discernible within it; an intensification of order 2 must have two such levels — that of individual segments and intensifications of order 1.) However, the *other* constituent of such an intensification need only be of order *n* - 1 *or less*. To take an example: the non-zero intensification which proves the superiority of *A* over *B* in the example on page 105 is of order 3. It consists of an *A*-to-*B* intensification of order 2 and a *B*-to-*A* intensification of order zero (i.e. the *A*-segment of intensity 1). The former of these consists in turn of an *A*-to-*B* intensification of order 1 and another *B*-to-*A* intensification of order zero (i.e. the *A*-segment of intensity 3). And the *A*-to-*B* intensification of order 1 consists, as it must, of an *A*-segment and a *B*-segment (the 5 and the 10 respectively).

We are now in a position to state the general principle NCP. It runs as follows:

NCP(S)

For any *X* and *Y*, if the following condition is satisfied, then *X* is negatively utilitarianly superior to *Y*:

There is a segmentation *S* of *X* and *Y*, such that all segments under *S* are of constant intensity and there is a set $\underline{\alpha}$ of *X*-to-*Y* intensifications satisfying the following conditions:

(A) Every *X*-segment under *S* occurs as an ultimate constituent of some member of $\underline{\alpha}$.

(B) No segment under *S* occurs as an ultimate constituent of more than one member of $\underline{\alpha}$.

(C) At least one member of $\underline{\alpha}$ is non-zero in magnitude.

5. The Collapse of the General Principle

We have seen that it is somewhat natural to accept the general principle — NCP(S) — as a sufficient condition for the negative utilitarian superiority of one situation over another. In this section, it is shown that we cannot in fact take this natural route. First of all, it is easy to show that NCP gives the wrong result in the T.I.C.C.. It will be remembered that in the T.I.C.C., situation *A* consists of one thousand minutes of suffering at level 1 and situation *B* consists of one minute at level 20. Let us now show that in this choice-situation, NCP prefers *B* to *A*. First of all, the segmentation *S* is simply to be a division of *A* and *B* into one-minute periods. Now what are the members of $\underline{\alpha}$ to be? $\underline{\alpha}$ can in fact be permitted to have just one member, which can be constructed as follows: combine the single minute of suffering in *B* with the first minute of suffering in *A* to form an intensification of order 1. This is an intensification from *A* to *B*, since the intensity of the *B*-segment (20) is greater than that of the *A*-segment (1). Its magnitude is 20 - 1 = 19. Now combine *this* intensification with the second minute of suffering in *A*, producing an intensification of order 2. (The reader can here appreciate the importance of the practice, defended at such length in previous sections, of allowing an intensification of order *n* to contain, as one of its immediate constituents, an intensification of order *less* than *n* - 1. In this case, an intensification of order 2 is made up of one intensification of order 1 and another of order zero.) And this intensification is *also* from *A* to *B*, since the magnitude of its component *A*-to-*B* intensification (19) is greater than that of its component *B*-to-*A* intensification (1). Its *own* magnitude is 19 - 1 = 18. Now if we combine this new *A*-to-*B* intensification with the *third* minute of suffering in *A*, we obviously get an *A*-to-*B* intensification of order 3 and of magnitude 17. Clearly this process can be continued for some time. But by the time we have involved our *twentieth* minute of suffering in *A*, the magnitude of our *A*-to-*B* intensification will only be zero. If this intensification is then combined with the twenty-first minute of suffering in *A*, the result will be an intensification, not from *A* to *B*, but from *B to A* (since 1 is greater than zero). This intensification can stand as the sole member of the set α . It is easy to see that, thus defined, α satisfies all the conditions laid down in NCP(S) (with *B* for 'X' and *A* for 'Y'). Since every minute of suffering in *B* (there is of course only one) occurs as an ultimate constituent of α's sole member, it satisfies Condition (*A*). Since α *has* only one member, it also satisfies Condition (*B*). And since the magnitude of α's single member is 1, α also satisfies Condition (C). Hence *B* is superior to *A* by NCP(S).

In preferring *B* to *A* in the T.I.C.C., NCP(S) is of course agreeing with the Principle of Total Utility. Thus it should not come as much of a surprise to the reader to learn that NCP(S) is in fact equivalent to the negative version of this principle. This is rigorously proven in an appendix. Right now, I want to address the question of what can be done about this awkward situation. On the one hand, the Principle of Total Utility seems completely unacceptable. When applied to the T.I.C.C., it requires us to prefer agony for a short period of time to trivial discomfort for a very long period of time, and this — we have argued — is completely wrong. And yet on the other hand, when we try to construct a principle which does, on the face of it, seem acceptable, we find that it collapses back into the Principle of Total Utility. Can this impasse be overcome?

It might be suggested that instead of assenting to all the members of the hierarchy of perfect correlative principles (and their logical product in the form of NCP) we should affirm only NCP0 and NCP1. It is impossible to derive the undesired result in the T.I.C.C. with the combination of these two principles. (It is easy to see that one needs NCP20 at least to get this result.[10]) Furthermore, we have seen that from NCP2 on, the principles start to bear a close resemblance to the Total Utility Principle. And in confining ourselves to NCP0 and NCP1, we would not even need to assign cardinal values to intensities, an operation which is thought by many to be problematic. (NCP0 does not require any quantitative comparisons. And NCP1 only requires an ordering-relation.) Can we then adopt a theory in which only NCP0 and NCP1 are accepted as valid sufficient conditions for negative utilitarian superiority?

I do not think that we can. The trouble with this course is that neither NCP0 nor NCP1 ever allow us to say that a situation *X* is superior to a situation *Y* when *X* contains some piece of suffering which is more intense than any occurring in *Y*. And yet we have already seen that we do sometimes want to say this. (See pp. 99-100.) Indeed this is one of the main purposes of NCP2. NCP2 applies, for example, in cases where, in choosing *X* over *Y*, we help one person and harm another, and the person harmed by the choice of *X* ends up in a worse state than the other would end up in through the choice of *Y*, and yet the choice of *X* is morally correct, since it makes a bigger *difference* for the person helped by it than for the person harmed by it. If we do not assent to NCP2, then we apparently cannot justify a decision of this kind.[11] But if NCP2 is to be accepted, why not NCP3, NCP4 and all the rest of them?

The only solution that I can see to this problem is to reject any firm distinction between acceptable and unacceptable principles. We should say, rather, that each perfect correlative principle has a *degree* of acceptability which depends on its position in the hierarchy. The higher the position of a principle is, the less acceptable it is i.e. the less certain and unproblematic are the superiorities which it generates.[12] More accurately, NCP0 and NCP1 are both fully acceptable principles, and the decline sets in with NCP2 and continues progressively as we ascend the hierarchy. But since NCP2 is only the first principle below the level of full acceptability, it is still highly reasonable. We have already pointed out that the lowest-order principle by means of which one can 'prove' that B is superior to A in the T.I.C.C. is NCP20. But this principle is of far too high an order to be taken seriously. When applied to the T.I.C.C., it involves the construction of an intensification of extremely high order. The latter is a highly complex, abstract entity, which seems to count for little in comparison with the concrete reality of the suffering in B. NCP20 does not just concern itself with actual intensities; it deals with differences between differences between differences . . . between intensities. Without claiming that differences are completely irrelevant (which would require us to reject even NCP2), we can claim that the more complex and abstract they are, the less relevant they are.

But it may be objected that we do not really have this option, precisely because the acceptance of every principle in the hierarchy seems, by a logical extension, to be necessitated by the acceptance of its predecessors. But the question is: can this 'logical extension' be construed as a logically compelling argument? Let us go back to the example that introduced NCP2 on page 99. We imagined a choice between suffering at intensity 1 for one minute and at intensity 9 for a further minute in situation *A*, and suffering for two minutes at intensity 8 in situation *B*. We construed the choice as being in effect equivalent to choosing between having a 1 instead of an 8 and having an 8 instead of a 9. Since 8 - 1 is much greater than 9 - 8, we decided that the former, and therefore situation *A*, should be chosen. In effect, this involves seeing the relation between *A* and *B* as being divisible into two distinct aspects. The first aspect involves a comparison between the 1 in *A* and the first 8 in *B*. If this aspect exhausted the relation between *A* and *B*, we would say that *A* was superior to *B* by NCP1. The second aspect involves a comparison between the 9 in *A* and the second 8 in *B*. If *this* aspect exhausted the relation between *A* and *B*, we would say that *B* was superior to *A* by NCP1. The idea that *A* has an *overall* superiority over *B* arises from the fact that the *degree* of NCP1-type superiority involved in the first aspect is greater than that involved in the second. Here we are comparing the magnitudes of *differences* between the intensities of segments, and not the intensities of segments themselves, as we do when we apply NCP1. But the belief that *A* is superior to *B* seems to arise out of a sort of *analogy* with NCP1-type reasoning. It is as if we were saying that the choice between *A* and *B* is like this choice: having a pain whose intensity is equal to the difference between 1 and 8 *or* having a feeling of discomfort whose intensity is equal to the difference between

8 and 9. In such a case, NCP1 would require us to choose the latter, and it is really by analogy with such a case that we decide that _A_ is superior to _B_ in _this_ case. But analogies can only form the basis of persuasive arguments, not logically compelling ones. Hence the NCP1-type reasoning which supports the preference for _A_ over _B_ makes it very _attractive_ to conclude that _A_ is actually superior, but it does not _necessitate_ doing so. Generalizing, the acceptance of NCP1 makes it very attractive to accept NCP2 also, but it does not necessitate doing so. By reasoning similar to the foregoing, one can draw analogies between cases in which NCP3 is applicable and certain kinds of cases in which NCP2 is applicable. But again these analogies only form the basis of a _persuasive_ argument from NCP2 to NCP3, not a logically compelling one. Since NCP2 is itself only supported by a persuasive argument from NCP1, the small doubt which already existed in the case of NCP2 is doubled when we come to NCP3. And obviously the doubt will continue to increase as we ascend the hierarchy of perfect principles. This is the reason then, I suggest, for the fact that the _complexity_ of intensifications diminishes their significance, and thus also the validity of the principles which invoke them.

6. The Imperfect Correlative Principles

We have already pointed out that in order to produce the intuitively wrong conclusion in the T.I.C.C. (viz. that situation *B* is superior), one would need to use a perfect principle of very high order, and such principles are very implausible. This is the reason for repudiating the thesis that *B* is superior in the T.I.C.C.. But how do we explain our judgement that *A* is actually superior? In order to do this, it is necessary to postulate a new principle. This principle is called 'INCPO' (for 'Imperfect Correlative Principle Zero'). It may be stated thus:

> *INCPO (S)*
>
> For any *X* and *Y*, if *X* contains some suffering, but all of it is of trivial intensity compared with some of the suffering in *Y*, then *X* is negatively utilitarianly superior to *Y*.

INCPO bears an obvious resemblance to NCPO. Like the latter, it does not require any one-to-one correlation of periods of suffering in *X* onto periods of suffering in *Y*. In the case of NCPO, this is because there *is* no suffering in *X*. In the case of INCPO, it is because the existence of just one period of suffering in *Y* in comparison with which *all* the suffering in *X* is trivial is sufficient for the superiority of *X*. We do not — in the manner of NCP1 — have to ensure that each *X*-segment is matched with its *own* *Y*-segment of equal duration and at least as great an intensity. Let us illustrate this by showing how INCPO applies in the T.I.C.C.. In situation *B*, we have suffering of intensity 20. And in situation *A*, all the suffering is of intensity 1, which we assume to be trivial in comparison with 20. It does not matter that the total duration of the suffering in *A* is so great. Indeed, we could increase its duration to any length we liked, and *A* would still be superior. Trivial pain multiplied over a long period of time is still trivial.[13] This is the fundamental rationale for INCPO.

INCPO is of course a much less clear-cut principle than NCPO. It contains a vague word — the word 'trivial'. Since the extension of this term does not have clear-cut boundaries, there are cases in which the triviality or non-triviality of some piece of suffering in comparison with some other is simply indeterminate. But it does not follow that the principle is useless. There are cases in which everyone will agree that some suffering is of trivial intensity, as well as cases in which everyone will agree that it is not.

We can imagine cases which are similar to the T.I.C.C., but too complex to be handled by INCPO. For these we need higher-order imperfect principles. Consider, for example, the choice represented by the following table:

SITUATION A	SITUATION B
A thousand minutes at level 15	A thousand minutes at level 14
	One minute at level 20

This case differs from the original T.I.C.C., since there is now substantial suffering in *A*. Nevertheless, it still seems reasonable to argue that *A* is superior, in that the 'improvement' from *A* to *B* (15 to 14) is trivial compared with the suffering of intensity 20 in B. This reasoning is generalized in INCP1:

INCP1(S)

For any *X* and *Y*, if the following condition is satisfied, then *X* is negatively utilitarianly superior to *Y*:

There is a segmentation *S* of *X* and *Y*, such that all the segments under *S* are of constant intensity and the following conditions are satisfied:

(A) There is a one-to-one mapping *M* from a subset of all *X*-segments under *S* onto a subset of all *Y*-segments under *S*, such that, for any *X*-segment σ , *M*(σ) is equal in duration to σ ,

AND, IN ADDITION:

(B) EITHER:

(B1) At least one of the pairs constituting *M* is a non-zero *X*-to-*Y* intensification,

OR:

(B2) there are *Y*-segments not correlated by *M*,

AND, IN ADDITION:

(C) although there are either non-zero *Y*-to-*X* intensifications amongst the pairs consti-

> tuting *M* or *X*-segments not correlated by *M*, they are all of trivial magnitude in comparison with some *X*-to-*Y* intensification in *M* or some *Y*-segment not correlated by *M*.

INCP1 is closely modelled on NCP1. The differences of course are the requirement that the domain of *M* only be a *subset* of the *X*-segments plus the addition of clause (C). Whereas NCP1 simply *excludes* *X*-segments not correlated by *M* as well as non-zero *Y*-to-*X* intensifications within *M* itself, INCP1 merely requires that they all be of trivial magnitude compared with some *X*-to-*Y* intensification in *M* or some *Y*-segment not correlated by *M*.

Applying INCP1 to the example on p. 119 above, the obvious course is to let *M* map each minute at level 15 in *A* onto a minute at level 14 in *B*. Clause (B2) is satisfied, in virtue of the extra minute at level 20 in *B*. And although *M* consists entirely of non-zero *B-to-A* intensifications, rather than *A*-to-*B* intensifications, they are all of trivial magnitude in comparison with this uncorrelated *B*-segment. Thus clause (C) is also satisfied. Hence *A* is superior to *B* by INCP1.

Note that although INCP1 is primarily linked to NCP1 — the perfect principle of the same order — it can also be seen to have some resemblance to the next perfect principle, NCP2. For as in the case of the latter, the magnitude of a *difference* between intensities — as opposed to merely the intensities themselves — can be relevant to its application. Thus in the previous example, we had to compare the <u>difference</u> between 15 and 14 with 20. Because of this similarity between the two principles, there are cases in which either could be applied indifferently to give the same result. Thus consider this modification of the previous example:

SITUATION A	SITUATION B
One minute at level 15	One minute at level 14
	One minute at level 20

In this case, we could just as well apply NCP2 as INCP1 to get the result that <u>A</u> is superior to <u>B</u>. Because of the reduction of the 15 and 14 from a duration of one thousand minutes to a duration of one minute, we can get a <u>one-to-one mapping</u> of all non-zero <u>B</u>-to-<u>A</u> intensifications onto <u>A</u>-to-<u>B</u> intensifications which, by the conditions of NCP2, renders <u>A</u> superior. But neither of the two principles NCP2 and INCP1 can replace the other. For there are cases — like the original version of this example — where INCP1 can be applied, but NCP2 cannot. And there are also cases where only NCP2 can be applied. (The example on p. 103 falls into this category, since a segment of intensity 3 cannot be considered 'trivial' in comparison with an intensification of magnitude 5.)

Clearly the process of formulating imperfect principles, paralleling the original perfect principles, can be continued indefinitely. For any $n > 0$, both NCP*n* and INCP*n* are characterized by the fact that the highest order of intensification that they correlate is $n - 1$. And the degree of validity of imperfect principles will diminish as we ascend the hierarchy, just as in the case of the perfect principles.[14]

Imperfect principles can contradict perfect ones. For example, in the T.I.C.C., INCP0 gives the result that *A* is superior to *B*, but NCP20 favours *B*. In this case, the correct course seems to be to adhere to INCP0, since the complex constructions of NCP20 are too far above the level of actual segments to have any real significance. But there are more problematic cases than this. Consider the following choice-situation:[15]

SITUATION A	SITUATION B
One minute at level 21	One minute at level 20
One minute at level 20	One minute at level 19
One minute at level 19	One minute at level 18
etc	etc
One minute at level 1	One minute at level 21

By matching each *A*-segment referred to on a given line with the *B*-segment referred to on the same line, we can show that *A* is superior to *B* by INCP1 (S). But the trouble is that *A* and B are equal in disvalue by NCP1(E). Every period of suffering in A is matched by an equivalent period of suffering in B and vice-versa. By choosing a different mapping we get a directly conflicting result.

Intuitions may differ on this, but it seems to me that the judgement of NCP1(E) ought to be preferred in this case. Once the particular mapping involved in its application is pointed out, I find it very difficult to take seriously the suggestion that either situation could be worse than the other in terms of degree of suffering. Indeed I would suggest in general that each perfect principle is more acceptable than its corresponding imperfect principle of the same order. The imperfect principle always involves some 'fudging' that the perfect principle does not. The satisfaction of the imperfect principle is itself a sort of approximation to the satisfaction of the corresponding perfect one, so if there is a different way in which the perfect principle can be exactly satisfied, this should prevail. However, I do not want to suggest that all conflicts resulting from different ways of applying the correlative theory admit of a clear resolution in favour of a particular manner of application. There may well be indeterminate cases.[16]

ENDNOTES

[1] The significance of the 'S' will appear presently.

[2] The significance of the 'S' and 'Version A' will

appear presently.

[3] This 'either . . . or' is of course intended to be inclusive, not exclusive.

[4] See Derek Parfit, *Reasons and Persons* (Oxford: Clarendon Press, 1984), Chapter 16, pp. 351-379.

[5] Parfit's remarks on p. 361 of *Reasons and Persons* serve to explain *why* they might be totally different. This is a consequence of the fact that the identity of people depends exactly on *when* they are conceived.

In my example, I have added the stipulation that the *q*-population is smaller. This is to make it more likely that the conditions of NCP1 would be satisfied. But it also has a practical plausibility, in that population-size and quality of life can obviously be inversely related.

[6] Of course, the discussion here only represents the *beginnings* of a utilitarian analysis. To develop it further, one would have to consider: (a) the effects of *p* and *q* during other periods of time besides *T*, and (b) the *positive* utilitarian effects of *p* and *q* i.e. their pleasure-producing effects.

[7] 'Involved in' means that the segment is either *identical* to *I* (or $M_2(I)$) or one of the segments of which *I* (or $M_2(I)$) is a pair.

[8] Note that there are no zero-intensity segments. As explained on page 91, a segmentation of *X* and *Y* is formed from a 'superlife' which contains only suffering.

[9] This is only a heuristic device. As pointed out in the note above, the theory proper does not bother to recognize segments of zero intensity.

[10] The application of NCP20 to this case would involve the construction of an *A*-to-*B* intensification of order 19 and magnitude 1, which would be matched by an equally intense *A*-segment, and still leave large numbers of other *A*-segments uncorrelated, thus showing *B* to be superior. This is simpler than the reasoning described earlier, which, since it involved correlating an intensification of order 20, would correspond to NCP21, rather than NCP20.

[11] In extreme cases, we could do so using the *imperfect* principle corresponding to NCP1. But we ought to be able to do so in *less* extreme cases as well. (This will be clarified in the next section.)

Some readers may feel that while NCP2 is not acceptable as it stands, a weaker version of it would be. It might be insisted that M_1 correlate only *contemporaneous* experiences that are had by a single person. That way one could argue that when *X* is superior to *Y* by NCP2, for any moment at which some person is worse off in *X* than in *Y*, there is a correlate moment at which he is worse off in *Y* than in *X* and to at least as great a degree. This perhaps makes the application of NCP2 more intuitive. The difficulty with it is that it makes the temporal order in which experiences occur a relevant feature of the utilitarian evaluation of situations, which it surely should not be. For on this view, *X* might be superior to *Y*, but not superior to some situation *Y'* which was identical to *Y*, except that the experiences in it occurred in a different order. This appears highly implausible.

[12] More of what is involved in attributing different degrees of validity to the principles is explained in Chapter Five, on page 133.

[13] It has been suggested to me that this might be true by definition and thus that INCP0 might be analytic. But while the first could well be the case, I do not think that the second follows from it. INCP0 states that, under the circumstances described, *X* *is preferable* to *Y* from a certain point of view. One is only *logically obliged* to accept this if one affirms suffering in general to be bad.

[14] Strictly speaking, however, there is a disanalogy here, in that in the case of the imperfect principles, the decline in validity would begin with INCP1, not INCP2, since the former is the first imperfect principle to permit comparisons of *differences* between intensities.

[15] The essence of which I owe to David Hitchcock.

[16] Another oddity in the relation of correlative superiority (its non-transitivity) is discussed in Appendix B (pp. 159-162).

CHAPTER FIVE

CONCLUDING CHAPTER

My purpose in this chapter is to round off the inquiry, firstly, by relating the results produced to the theoretical structure described in Chapter One, involving the recognition of a multiplicity of types of _prima facie_ superiority, combining to determine overall superiorities, and secondly, by making some concluding remarks concerning the thesis as a whole.

1. The Balancing of _Prima Facie_ Superiorities

We have seen that the concept of 'maximizing happiness' involves _two_ distinct kinds of _prima facie_ superiority — one positive, involving pleasure; and the other negative, involving pain or suffering. But we have also seen that each of these kinds of superiority themselves involve a multiplicity of principles which sometimes conflict, and which then have to be balanced against one another. This raises the possibility that even '_X_ involves less suffering than _Y_' does not represent a _single_ form of _prima facie_ superiority, but that there is a distinct form represented by each of the negative correlative principles. Such a view would make the existence of conflicts between the principles appear more acceptable. It is easy to understand how *X* can be superior to *Y* in one respect, but *Y* superior to *X* in another.

But in fact we cannot adopt this view. Consider again the example introduced at the end of the previous chapter:

SITUATION A	SITUATION B
One minute at level 21	One minute at level 20
One minute at level 20	One minute at level 19
One minute at level 19	One minute at level 18
etc	etc
One minute at level 1	One minute at level 21

It will be remembered that by NCP1(E), *A* and *B* are equal in disvalue, but by INCP1(S), *A* is superior. I have maintained that the former judgement is the correct one. Neither situation is any worse, from a negative utilitarian point of view, than the other. This seems to preclude regarding NCP1 and INCP1 as representing two different forms of *prima facie* superiority. Remember that for any *X* and *Y*, if *X* is *prima facie* preferable to *Y*, and *Y* is in no way *prima facie* preferable to *X*, then, by what I have called the '*Ceteris Paribus* Axiom' (page 3), *X* is preferable *overall* to *Y*. But if *X* is *prima facie* preferable to *Y*, and *Y* is also *prima facie* preferable to *X*, we cannot so easily conclude that *X* has an overall superiority over *Y*. Any respect in which *Y* is superior to *X* constitutes an obstacle to concluding that

X has overall superiority on the basis of *X*'s *prima facie* superiority. In this case, we have to *weigh* the respects in which *X* is better than *Y* against the respects in which *Y* is better than *X*. But obviously no such obstacle would be created by the knowledge that *X* and *Y* were *equal* in some respect. Indeed such knowledge would simply call for the situations to be compared on some other basis — e.g. the very basis on which *X* is being judged *prima facie* superior. Thus if NCP1 and INCP1 simply represented different bases of comparison in this sense, we would not expect the fact that *A* and *B* were equal in terms of NCP1 to constitute an obstacle to following INCP1's judgement that *A* is superior. Indeed this fact would have to be considered as *strengthening* the significance of INCP1's judgement, since it entails that *B* is not superior to *A* by NCP1 — i.e. that a certain kind of possible *prima facie* superiority of *B* over *A* that would have conflicted with the *prima facie* superiority of *A* over *B* does not in fact obtain. But this is not the case. Not only does the equality of *A* and *B* constitute an obstacle to our accepting the judgement of INCP1 (S) — it actually seems to *override* the latter and lead us to conclude that there is nothing to choose between *A* and *B*.

We cannot therefore say that all the correlative principles represent different forms of *prima facie* superiority. Perhaps this thesis is very implausible anyway in the case of NCP1 and INCP1, since, as was pointed out in the previous chapter, INCP1 is merely a less 'rigorous' version of NCP1. One might not want to say, then, that a perfect principle represented a different kind of superiority from its corresponding imperfect principle of the same order. But one might nevertheless wish to maintain that each correlative principle — whether in its perfect or imperfect form — represents a different form of superiority from all the other principles of different orders. But clearly this would not work any better. If, for example, one found a case in which _A_ and _B_ were equal by NCP1(E), but _A_ was superior to _B_ by INCP2(S), one would, contrary to what would be expected on this view, conclude that neither _A_ nor _B_ was any worse than the other.

But if we do not say that the different correlative principles represent distinct (albeit related) forms of _prima facie_ superiority, what _should_ we say about them? The answer is that we should consider them to be different (and sometimes conflicting) expressions of a single concept — that of one situation's being superior to another when considered by someone in relation to his aversion to (general) suffering. The conflicts are to be explained by the fact that it sometimes happens that, when looked at from one point of view (in terms of a particular principle and a particular mapping or set of mappings), _X_ seems to exhibit this superiority to _Y_, while from another, it appears not to. This is an admittedly less comfortable position than the idea of the conflicts as arising through a plurality of distinct forms of superiority, but I do not see it as genuinely unacceptable.

However, it is clear that when we come to the relation between positive and negative utilitarian superiority, two distinct kinds of _prima facie_ superiority _should_ be recognized. There is no difficulty in this case parallel to that involved in the previous one. When _X_ involves less suffering than _Y_, _X_'s being equal to _Y_ with respect to its amount of pleasure does not constitute any obstacle to concluding that _X_ has overall utilitarian superiority over _Y_ — on the contrary, it confirms the latter judgement. However, we clearly will have cases in which positive and negative superiority conflict. Furthermore, we will also have cases in which either positive or negative superiority (or both) conflict with a judgement made on the basis of some non-utilitarian form of _prima facie_ superiority. Sometimes such conflicts are irresolvable. But there clearly are cases in which we can judge a certain _prima facie_ superiority to be more weighty and significant than some other with which it conflicts, and thus that the former should override the latter. The ability to do this depends upon a concept which we have not so far said anything about — that of _degree_ of _prima facie_ preferability. We do not just speak of one situation's being preferable to another (_prima facie_ or _simpliciter_). We often also speak of its being preferable *to a greater or lesser degree*. This raises the question of what principle or principles such judgements are based on.

In certain cases, comparisons of degrees of *prima facie* preferability are very easy to make. Take, for example, the following two cases (the 'levels' are, as usual, intensities of suffering):

(1)	SITUATION A	SITUATION B
	One minute at level 5	One minute at level 10
(2)	SITUATION A	SITUATION B
	One minute at level 10	One minute at level 15

It is easy to see that in both cases, *A* is superior to *B* (by NCP1) *and to the same degree*. The degree of superiority can simply be identified with the numerical difference between the intensities on each side, and this is 5 in both cases. (Such comparisons are of course the basis for the application of NCP2.) We can sometimes even make clear judgements of comparison in cases where the superiorities are based on different correlative principles. Consider:

(3)	SITUATION A	SITUATION B
	One minute at level 10	One minute at level 20
	One minute at level 5	

In this case, *A* is superior to *B* by NCP2. And we can identify the degree of preferability with the magnitude of the intensification of order 2 which is involved in its application viz. 5. (We can either map the 10 onto the 20 and map the 5 onto the resulting pair, or we can map the 10 onto the difference between 5 and 20. Either way, the magnitude of the higher-order intensification is 5.) Thus we can see that Case (3) involves the same degree of preferability at the level of NCP2 as do cases (1) and (2) at the level of NCP1. However, it should be noted that this identical degree of difference will not have the same significance at the two levels. NCP2 is a less acceptable principle than NCP1. This means that a given degree of preferability obtaining at the NCP2-level will not have the same weight — that is, will not be capable of outweighing as large a range of possible conflicting *prima facie* superiorities — as the same degree obtaining at the NCP1-level. This is really what the attribution of different degrees of validity to the different correlative principles actually amounts to in practice. Note that this approach entails that some applications of NCP2 are actually *more* significant than some applications of NCP1. For example, a very large NCP2-difference should have much more weight than a minute NCP1-difference.

But this still leaves the question of how 'degree of superiority' can be measured in general. Here we should remind ourselves of the fundamental difference between the present method of judging superiorities and the method of absolute disvalues.[1] In the case of the latter, the determination of degrees of superiority is easy. The degree to which X is superior to Y is simply the numerical difference between the absolute disvalue of X and the absolute disvalue of Y. But since on the present method we do not attribute absolute disvalues to situations, but instead make the relation of 'having less suffering' itself the fundamental one, this course is not available to us. Is there any method of defining a measure of degrees of superiority in terms of the correlative principles which we have adopted for determining the existence of these superiorities in the first place? In the examples described earlier, the equality in degree of superiority seemed to depend on the equality of the magnitude of the 'final' non-zero X-to-Y intensifications that proved the superiority of X over Y. But what if we have more than one such intensification? Can we perhaps just sum their magnitudes? More accurately, the suggestion might be this: sum the results of multiplying the magnitude of each 'effective' intensification by the single duration shared by all its ultimate constituents, where an 'effective' intensification is either a non-zero X-to-Y intensification from amongst the pairs constituting the final mapping in the application of the correlative principle in question, or a non-zero X-to-Y intensification left uncorrelated by that mapping.[2] But there are serious problems with this proposal. In the first place, as it stands, it can only handle the correlative principles of order greater than zero. If one were to extend the method so as to cover NCPO and INCPO, one would apparently have to say that whenever _X_ was superior to _Y_ by either of these two principles, the degree of superiority was equal to the total disutility of _Y_. But if any situation involving suffering is negatively inferior to one which does _not_ involve any suffering to a degree equal to the total disutility of the former, this is surely tantamount to saying that every situation has an absolute disvalue equal to its total disutility — i.e. it is tantamount to (Negative) Total Utilitarianism, which we have already rejected.[3] Secondly, as might therefore be expected, the proposal does not even give the right results when applied to the correlative principles of order greater than zero. Suppose _B_ is identical to _A_, except for the addition of 20 extra one-minute periods, each of intensity 1. Then the degree of superiority of _A_ over _B_ would, on the present proposal, be exactly the same as if the difference were just a matter of _one_ extra minute at intensity 20. But of course, it is much less.

The method just considered attempts to describe a _cardinal measure_ for degrees of superiority. And one of the arguments used against it can be used to show that _no_ cardinal measure is possible. For suppose some particular measure is adopted. Then it must specify cardinal values for the degree of inferiority of different kinds of situations involving suffering over other kinds of situations, including those _without_ any suffering. And in the latter case, these degrees of inferiority will amount to absolute disvalues for the inferior situations. In other words, it appears to be impossible to adopt a cardinal measure for degrees of inferiority, once the idea of assigning absolute disvalues to situations themselves has been abandoned.

Nevertheless, presumably some sort of theory of the conditions under which one instance of utilitarian superiority could be said to be superior to another can be produced. And one would expect it to be structurally similar to the conglomeration of principles which we have suggested for determining when one _situation_ is utilitarianly superior to another. I shall not pursue the issue further here. It would be a good topic for future research.

Let us now return to the problem that initiated this discussion of degrees of preferability in the first place — the balancing of conflicting _prima facie_ superiorities. One important point to notice is the high degree of indeterminacy in this matter — an indeterminacy that would still exist even if we did have a satisfactory theory of degrees of preferability. Let us consider a very simple example of the type of conflict which is of most significance to the utilitarian — that between positive and negative utilitarian superiority. Suppose _A_ contains only hedonically neutral experiences, while _B_ contains one minute of moderate-intensity pleasure as well as one minute of suffering. Clearly _A_ is negatively superior to _B_, but positively inferior to _B_ (because it lacks the pleasure that _B_ contains). But either situation may be superior _overall_ to the other, depending on the relative weight of the positive superiority of _B_ over _A_ and the negative superiority of _A_ over _B_. The question is: when does one outweigh the other? How would the intensity of the suffering in _B_ have to compare with the intensity of the pleasure for the positive superiority to outweigh the negative one and vice-versa? I have already argued in Chapter Two that although it is possible to attach a sense to the suggestion that a pleasurable experience and an unpleasant one are equally intense, it is unreasonable to require that such judgements should govern the question of normative comparison. In the absence of any clear standard for making such comparisons, a double indeterminacy is created. One is specific to any individual utilitarian. If, as the intensity of the pleasure in _B_ remains constant, we imagine the intensity of the suffering gradually increased, starting at its 'minimally objectionable' level, we can see that to begin with, the pleasure will outweigh the suffering i.e. the positive superiority of _B_ over _A_ will outweigh the negative superiority of _A_ over _B_, resulting in a net superiority for _B_. Eventually, however, the intensity of the suffering will be sufficiently great to reverse this judgement. But there will be a broad intermediate range in which this utilitarian will simply be unable to decide whether the presence of the pleasure in _B_ counts for more than the absence of the suffering in _A_ or vice-versa i.e. a range in which neither superiority outweighs the other. This indeterminacy is then compounded if we go beyond the individual utilitarian, and consider utilitarians _generally_. For we shall then find disagreement even about the location of the points where the (original) indeterminacy ceases and one superiority outweighs the other.[4] And this is only a very simple example. No doubt if we varied _durations_ as well as intensities, even greater uncertainty would be produced.

These facts are frustrating from a theoretical point of view. But from the standpoint of the practical application of utilitarianism, they are less important than might be thought. The utilitarian should not be spending too much time wondering whether various pleasures are worth their cost in terms of suffering. For this would be to assume that there is a rigid unalterable connection between the pleasures that people experience and the particular costs by which, under _present_ conditions, they are accompanied. He should rather be concentrating on finding ways of obtaining _low-cost_ pleasures. To try to prevent people from pursuing certain pleasures because one judges the cost in suffering to be too high is likely to be a much less effective policy than attempting to find for them ways of obtaining the same pleasures at a lower cost. In general, judgements of utilitarian superiority which are hard to make and which, when made, are even harder to defend, can often be _evaded_ by choosing more radical policies that will lead to situations that are clearly better, from a utilitarian point of view, than any of those previously considered.

Besides the weighing of positive and negative superiority in order to determine overall utilitarian superiority, we must also consider the weighing of utilitarian superiority with non-utilitarian forms of moral _prima facie_ superiority. It is important to note here that a person can recognize that other moral considerations besides utilitarian ones have weight, while still basically counting as a 'utilitarian'. For example, when two situations involve equal amounts of pleasure and suffering, one may nevertheless be morally superior to the other in possessing a better apportionment of pleasure and suffering to desert.[5] Such a consideration might even _outweigh_, in some cases, a definite utilitarian superiority — for example, if the latter involved only a very trivial difference in degree of suffering over a short period of time. What, then, makes someone a 'utilitarian'? It seems reasonable to suggest that to be considered a utilitarian, one must regard a sufficient degree of utilitarian superiority as absolutely overriding in relation to other possible moral considerations.[6] For example, if one situation is negatively superior to another by NCP1, in such a way that one can produce an NCP1-type mapping that leaves uncorrelated in the second situation many high-intensity segments covering a very long total duration, and the second situation is not positively superior to the first, then the utilitarian ought definitely to consider the first situation to be superior from the moral standpoint, however it might compare to the second in other non-utilitarian ways. Such a restriction, although weaker than one might expect, will have a surprisingly large effect on the general approach to moral and social problems of one who accepts it. In particular, it will mean that he will want to ensure that as far as possible, people are caused to have those attitudes and habits of mind that are most productive of happiness. For suppose someone advocates certain child-rearing practices or educational methods which, though perhaps in accordance with traditional morality, are likely to produce less happy people than other possible methods. Then the utilitarian will reject them, for the difference, in utilitarian terms, between their consequences and those of the superior methods would easily be serious enough to override any other considerations, since people's basic dispositions and attitudes obviously have such a profound <u>long-term</u> effect on their happiness. In fact a utilitarian will want to encourage active research into determining what the best child-rearing and educational practices would be from the perspective of happiness-maximization. However, it is important to note that, as several writers have pointed out, the best dispositions to instill — best, that is, even in utilitarian terms — are not necessarily purely utilitarian ones.[7] It may turn out to be the case that more happiness can be produced in the long run by instilling in people some non-utilitarian attitudes. The utilitarian could even apply this to his own case. He knows that he has some (non-instrumental) _prima facie_ moral principles that are not utilitarian. But he could, without incoherence, think it better, _even in utilitarian terms_, that he and others should have these principles, rather than having only utilitarian ones. (This is not to deny of course that he could also think that some of his attitudes were utilitarianly counter-productive and that it would be better if he did not have them.)

2. Concluding Remarks

I wish to end by reflecting on what I think I have achieved in this thesis, and what still needs to be done. Utilitarianism is an attractive theory, but it is also one which has met with a considerable amount of criticism. The criticisms can be classified according to whether they concern the attitudinal part or the behavioural part of utilitarian theory i.e., roughly speaking, according to whether they concern the utilitarian's theory of good and bad states of affairs or his theory of right and wrong conduct. The basic criticism with respect to the attitudinal part is that the utilitarian's concept of 'amount of happiness', understood in an 'impersonal' way, either makes no sense at all, or else produces counter-intuitive results. This is the criticism I believe myself to have answered, at least in part. In effect, what I have said is that the counter-intuitive results can only be derived from theories — like Total Utilitarianism and Average Utilitarianism — which impose a structure that is overly simple in relation to what our intuitions actually seem to demand. I have instead produced a theory which I believe does not suffer from this defect. Of course, while the conventional theories might be criticized for being too simple, my theory might be criticized for being too complex and inelegant, especially in its admission of objective indeterminacies, and its rejection of any firm distinction between valid and invalid principles. But this is perhaps the price to be paid for sticking close to intuition. Our intuitions themselves seem to be complex and perhaps even 'inelegant' in their structure. I therefore do not believe that the theory should be rejected on this count. I have, however, conceded that it needs further development, particularly in the clarification of the concept of 'degree of utilitarian superiority'.

With regard to the behavioural component, the most common type of criticism takes the form of suggesting that utilitarianism diverges too much in its consequences from common-sense morality. Assuming that one does want to be regarded as a 'utilitarian' in some 'all-encompassing' sense, and not accept completely independent non-utilitarian criteria in a pluralistic spirit (which is also perfectly possible), one can answer this criticism in a manner which has already been anticipated to some extent. One can say that utilitarianism is to be applied in an *indirect* way, so that our having many of the common-sense principles of conduct that we do is seen to be something which is in accordance with utilitarianism itself, in that it can reasonably be argued to be more conducive to happiness than our following principles of conduct that are strictly utilitarian. But there is clearly much more work to be done on the behavioural component of utilitarian theory. In particular, as I suggested in my opening chapter, the further development of the concept of *co-operative* utilitarian action has an important role to play here.

ENDNOTES

[1] See pp. 80-83.

[2] In the case of NCP1, these two types of intensification correspond, respectively, to clauses (B1) and (B2) (see p. 92), and in the case of NCP2, to clauses (B4.1) and (B4.2) (see p. 102).

In the case of the perfect principles the proposal is equivalent to identifying the degree of superiority with the difference in total disutility — but not of course, in the case of the imperfect ones.

[3] The argument could also be expressed by saying that if *X* is worse than a suffering-free situation to degree *n*, and *Y* is worse than a suffering-free situation to degree *m*, and $n < m$, then *Y* must be worse than *X*, and thus — on the

proposal being considered — *Y* would have to be worse than *X* whenever its total disutility was higher. The rejection of this argument seems implausible, even if it is not actually incoherent.

[4] Of course, the possibility of irresolvable disagreement was already recognized in Chapter Two with respect to the actual assignment of intensities to the experiences. (See p. 48.)

[5] As Ross says, 'if we compare two imaginary states of the universe, alike in the total amounts of virtue and vice and of pleasure and pain present in the two, but in one of which the virtuous were all happy and the vicious miserable, while in the other the virtuous were miserable and the vicious happy, very few people would hesitate to say that the first was a much better state of the universe than the second.' (*The Right and the Good* (Oxford: Clarendon Press, 1930), p. 138.) Of course, I am, like Ross, talking about a *non-instrumental* preference. The preference could be *instrumentally* utilitarian, in which case it would present no problem. But it *can* also be non-instrumental — imagine that Ross's comparison is between two *entire* universes instead of between two *states* of a universe.

[6] The acceptance of such a position is not essential, however, to the main thrust of this thesis. The utilitarian principles suggested can be used in a system in which they are never considered as absolutely overriding. They may be accepted as having *some* weight by people who could not be regarded as 'basically utilitarian'.

[7] See, for example, R. M. Hare, *Moral Thinking: its Levels, Method and Point* (Oxford: Clarendon Press, 1981), Sec. 8.3, pp. 135-137, and Derek Parfit's discussion of how consequentialism can be 'indirectly self-defeating' in *Reasons and Persons* (Oxford: Clarendon Press, 1984), pp. 24-28.

APPENDIX A

PROOF OF THE EQUIVALENCE THESIS

In this appendix, I am going to provide a rigorous proof of the claim made in Section Five of Chapter Four that the general principle of negative perfect correlative superiority — NCP (S) — is logically equivalent to the principle of minimizing total disutility. It is this claim that I am referring to as the 'Equivalence Thesis'. It may be split up into two parts, which I shall refer to as 'E1' and 'E2':

E1: NCP (S) entails the principle of minimizing total disutility.

E2: The principle of minimizing total disutility entails NCP (S).

I shall prove E2 first. E2 states that if we accept that any situation is negatively superior to any other than which it has a lower total disutility, then we are logically obliged to accept NCP (S). This will immediately follow if we can show that the conditions laid down in NCP (S) guarantee that *X* has a lower total disutility than *Y*.

If NCP (S) is satisfied, then there is a segmentation *S* of *X* and *Y* involving only constant-intensity segments from which one can construct a set $\underline{\alpha}$ of *X*-to-*Y* intensifications satisfying the appropriate conditions.[1] Now either there is suffering in *X* or there is not. If there is not, then *α* consists entirely of *Y*-segments (Condition (A) being vacuously satisfied)[2] and their existence obviously entails that the sum of suffering in *X* is less than that in *Y*. If, on the other hand, there is suffering in *X*, then by Condition (A), at least one member of *α* is not merely a *Y*-segment. Corresponding to each such member, there will be a statement of the form '$N \leqslant M$', which is the justification for regarding the member as an *X*-to-*Y* intensification. Since *N* is the magnitude of a *Y*-to-*X* intensification, it is calculated through *adding* one or more *X*-segment-magnitudes and (where this intensification is of order greater than zero) *subtracting* one or more *Y*-segment-magnitudes. And since *M* is the magnitude of an *X*-to-*Y* intensification, it is calculated through *adding* one or more *Y*-segment-magnitudes and (where this intensification is of order greater than zero) *subtracting* one or more *X*-segment-magnitudes. In the case of each of these statements, we are mathematically entitled to transfer any *Y*-segment-magnitudes subtracted from the left-hand-side onto the right-hand-side (where they will be added), and we are similarly entitled to transfer any *X*-segment-magnitudes subtracted from the right-hand-side onto the left-hand-side (where again they will be added). This will produce, in each case, a statement of the form '$N' \leqslant M'$' in which *N*' is a sum of *X*-segment-magnitudes and *M*' is a sum of *Y*-segment-magnitudes. Now by Condition (A) of NCP(S), every *X*-segment occurs as an ultimate constituent of some member of $\underline{\alpha}$. Therefore every *X*-segment-magnitude is covered in at least one of these statements. In virtue of Condition *(B)*, no segment (from *X* or *Y*) occurs as an ultimate constituent of *more* than one member of $\underline{\alpha}$. This means that no *X*-segment or *Y*-segment is represented in *more* than one of these statements. Furthermore, clause 2(c) of the definition of an intensification, which prevents segments from being repeated within a single intensification, entails that no segment is represented more than once *within* a statement. Next, in virtue of clause *2(d)* of this definition, all segments represented within a given statement are of equal duration. Finally, in virtue of clause (C) of NCP(S), which requires that at least one member of $\underline{\alpha}$ be non-zero in magnitude, either there is a *Y*-segment not covered in any of the statements or, for at least one of the statements, the '$\leqslant$' can be replaced by a '$<$' without change of truth-value.[3] It is easy to see that the conjunction of all these facts entails, by some very elementary algebraic manipulation, that the sum of the results of multiplying the intensity of each *X*-segment by its duration is less than the sum of the results of multiplying the intensity of each *Y*-segment by *its* duration i.e. the total disutility of *X* is less than that of *Y*. This completes the proof of E2.

We now turn to the proof of E1, which states that NCP(S) entails the principle of minimizing total disutility.

This is proven by showing that, for any $\underline{X}$ and $\underline{Y}$, if $\underline{X}$ has a lower total disutility than $\underline{Y}$, then the conditions laid down in NCP(S) must be satisfied.

First of all, we distinguish two situations in which the total disutility of $\underline{X}$ is lower than that of $\underline{Y}$:
Case (a): There is no suffering in $\underline{X}$ and some suffering in $\underline{Y}$.
Case (b): There is suffering in both situations.
Case (a) is simple to deal with. All that would have to be done to show that NCP(S) was satisfied would be to come up with any segmentation of $\underline{Y}$ in which all the segments were of constant intensity, and let $\underline{\alpha}$ be any subset of these segments.[4]

Case (b) is the difficult one. Consider any segmentation $\underline{S}$ of $\underline{X}$ and $\underline{Y}$, such that all the segments under $\underline{S}$ are of constant intensity and all of exactly equal duration.[5] Let the number of $\underline{X}$-segments be $\underline{n}$ and the number of $\underline{Y}$-segments $\underline{r}$. Let all the $\underline{X}$-segments be numbered from 1 through $\underline{n}$ and all the $\underline{Y}$-segments from 1 through $\underline{r}$, and let the variable '$\underline{x}_i$' represent the $\underline{X}$-segments and the variable '$\underline{y}_j$' the $\underline{Y}$-segments. The expression '$\mu(\ )$' can be used to represent the intensity of a segment, or, more generally, the magnitude of an intensification. Given this symbolic apparatus, the statement that the total disutility of $\underline{X}$ is lower than that of $\underline{Y}$ can be expressed as follows:

$$\sum_{i=1}^{n} \mu(\underline{x}_i) < \sum_{j=1}^{r} \mu(\underline{y}_j)$$

I am going to prove that if this statement is true, then we can always construct, from the segments $\underline{x}_1 \ldots \underline{x}_n$ and at least some of $\underline{y}_1 \ldots \underline{y}_r$, a set $\underline{\alpha}$ of $\underline{X}$-to-$\underline{Y}$ intensifications satisfying the conditions laid down in NCP(S). In fact I am going to prove something even stronger than this viz. that we can always construct a <u>one-membered</u> set of this kind. In all the different cases discussed, this one member of $\underline{\alpha}$ will be referred to as the 'final' $\underline{X}$-to-$\underline{Y}$ intensification for that case. ('Final' because it is usually the end-result of a process of construction.)

We distinguish cases primarily according to the value of $\underline{r}$ i.e. according to the number of $\underline{Y}$-segments there are. In case (1), there is just one $\underline{Y}$-segment and an undetermined number of $\underline{X}$-segments. In case (2), there are two $\underline{Y}$-segments with an undetermined number of $\underline{X}$-segments; in case (3), three $\underline{Y}$-segments and so on.

The proof is inductive in nature. The basis of the induction is the proof that the thesis holds in case (1). The induction step is a proof that for any $\underline{r} > 1$, if the thesis holds in case $\underline{r} - 1$, then it holds in case $\underline{r}$.

<u>Case 1 (basis)</u>

In this case, there is only one $\underline{Y}$-segment $\underline{y}_1$. The assertion that the total disutility of $\underline{X}$ is less than that of $\underline{Y}$ amounts to the following:

(1) $$\sum_{i=1}^{n} \mu(\underline{x}_i) < \mu(\underline{y}_1)$$ [6]

The construction of an appropriate set $\underline{\alpha}$ is very simple in this case. The one member of $\underline{\alpha}$ is built up as follows: we start with the zero-order *Y*-to-*X* intensification $\underline{x}_1$. This combines with the zero-order *X*-to-*Y* intensification $\underline{y}_1$ to produce a new *X*-to-*Y* intensification of order *one*. We can be sure that this combination *is* an *X*-to-*Y* intensification, since $\mu(\underline{x}_1)$ must be less than $\mu(\underline{y}_1)$, for, in virtue of (1) above, the sum of *all* the $\mu(\underline{x}_i)$s is less than $\mu(\underline{y}_1)$. We now combine this new *X*-to-*Y* intensification with the *Y*-to-*X* intensification $\underline{x}_2$. The result must again be an *X*-to-*Y* intensification since, again in virtue of (1), $\mu(\underline{x}_2) < \mu(\underline{y}_1) - \mu(\underline{x}_1)$. We then combine *this* *X*-to-*Y* intensification with $\underline{x}_3$ and so on until we have exhausted all the $\underline{x}_i$s. Call the final *X*-to-*Y* intensification thus produced '*I*'. (Its magnitude is $\mu(y_1) - \sum_{i=1}^{n} \mu(\underline{x}_i)$.) *I* is the sole member of a set of intensifications satisfying the three conditions laid down in NCP(S). Since all the *X*-segments figure as ultimate constituents of *I*, Condition (A) is certainly satisfied. And since *I* is the only member of the set, (B) is also satisfied. Finally, since $\mu(\underline{x}_n)$ must be *less* than $\mu(\underline{y}_1) - \sum_{i=1}^{n-1} \mu(\underline{x}_i)$ and cannot be equal to it, *I* is non-zero in magnitude and thus Condition (C) is satisfied too.

Case r (Induction step)

We now consider the case of any undetermined number

$\underline{r}$ of $\underline{Y}$-segments (greater than 1). We assume that the thesis holds in case $\underline{r} - 1$, and we want to use this fact to prove that it also holds in case $\underline{r}$. This case is defined by statement ($\underline{r}$):

$$(\underline{r}) \quad \sum_{i=1}^{n} \mu(\underline{x}_i) < \sum_{j=1}^{r} \mu(\underline{y}_j)$$

Case $\underline{r}$ may be split up into two possible sub-cases:

($\underline{r}$.1) For some $\underline{y}_j$, $\sum_{i=1}^{n} \mu(\underline{x}_i) < \mu(\underline{y}_j)$

($\underline{r}$.2) For every $\underline{y}_j$, $\sum_{i=1}^{n} \mu(\underline{x}_i) \geq \mu(\underline{y}_j)$

We shall deal with each in turn.

<u>r.1</u>

For some $\underline{y}_j$, $\mu(\underline{y}_j)$ is greater than the sum of all the $\mu(\underline{x}_i)$s. Take any arbitrary such $\underline{y}_j$ and call it '$\underline{y}_a$'. By the reasoning of case (1), there is an intensification, whose magnitude is $\mu(\underline{y}_a) - \sum_{i=1}^{n} \mu(\underline{x}_i)$, which can stand as the final $\underline{X}$-to-$\underline{Y}$ intensification.

<u>r.2</u>

We have:

$$\sum_{i=1}^{n} \mu(\underline{x}_i) \geq \mu(\underline{y}_1)$$

Transferring all but the last $\mu(\underline{x}_i)$ to the right-hand-side of the statement, we get:

$$\text{(i)} \quad \mu(\underline{x}_n) \geq \mu(\underline{y}_1) - \sum_{i=1}^{n-1} \mu(\underline{x}_i)$$ [7]

Now consider $\underline{x}_1$. It may be that its magnitude alone is greater than or equal to that of $\underline{y}_1$. If so, then $\mu(\underline{x}_1)-\mu(\underline{y}_1)$ is the magnitude of a *Y*-to-*X* intensification. (This intensification will be used to construct the final *X*-to-*Y* intensification.)

If, on the other hand, $\mu(\underline{x}_1)$ is *less* than $\mu(\underline{y}_1)$, then consider $\mu(\underline{x}_2)$. Compare it with $\mu(\underline{y}_1)-\mu(\underline{x}_1)$. Is it greater than or equal to the latter quantity or not? If it is, then $\mu(\underline{x}_2)-(\mu(\underline{y}_1)-\mu(\underline{x}_1))$ is the magnitude of a *Y*-to-*X* intensification. (This intensification will be used to construct the final *X*-to-*Y* intensification.) The reasoning to justify this runs as follows: by hypothesis, $\mu(\underline{x}_1)$ is less than $\mu(\underline{y}_1)$. Thus $\mu(\underline{y}_1)-\mu(\underline{x}_1)$ is the magnitude of an *X*-to-*Y* intensification. But also by hypothesis, $\mu(\underline{y}_1)-\mu(\underline{x}_1)$ is less than or equal to $\mu(\underline{x}_2)$. Since $\underline{x}_2$ is a *Y*-to-*X* intensification, it follows that $\mu(\underline{x}_2)-(\mu(\underline{y}_1)-\mu(\underline{x}_1))$ must be the magnitude of a *Y*-to-*X* intensification.

If, on the other hand, $\mu(\underline{x}_2)$ is *less* than $\mu(\underline{y}_1)-\mu(\underline{x}_1)$, then consider $\mu(\underline{x}_3)$. Compare it with $\mu(\underline{y}_1)-\mu(\underline{x}_1)-\mu(\underline{x}_2)$. Is it greater than or equal to the latter quantity or not? If it is, then by reasoning similar to that of the previous case, $\mu(\underline{x}_3)-(\mu(\underline{y}_1)-\mu(\underline{x}_1)-\mu(\underline{x}_2))$ must be the magnitude of a *Y*-to-*X* intensification. (This intensification will be used to construct the final *X*-to-*Y* intensification.) If it is not, then consider $\mu(\underline{x}_4)$, asking whether or not the latter is greater than or equal to $\mu(y_1)$ minus all the previous $\mu(x_i)$s. Because of the truth of (i) above, we can be sure that we will eventually hit on an x_i of which this is true, whether it be x_n or a previous one. Call the *first* one we hit on 'x_m'. Its intensity minus the result of subtracting all the previous $\mu(x_i)$s from $\mu(y_1)$ is the magnitude of the required *Y*-to-*X* intensification. (Note that it is crucial that we use the *first* x_i. For in order to be sure that $\mu(y_1)$ minus all the previous $\mu(x_i)$s is an *X*-to-*Y* intensification, we use the fact that each of these previous $\mu(x_i)$ s is *not* greater than or equal to $\mu(y_1)$ minus all the $\mu(x_i)$ s prior to *them*.)

There are now two possibilities to consider. Either:

(*r*.21) $x_m = x_n$

or

(*r*.22) $x_m \neq x_n$

r.21

$x_m = x_n$

From (*r*) above we can derive the following:

(ii) $\mu(x_n) - \left(\mu(y_1) - \sum_{i=1}^{n-1} \mu(x_i)\right) < \sum_{j=2}^{r} \mu(y_j)$ [8]

Since $x_m = x_n$, we know from previous reasoning concerning x_m that the left-hand-side of (ii) must be the magnitude of a *Y*-to-*X* intensification. Now consider $\mu(y_2)$. Either the left-hand-side of (ii) is less than $\mu(\underline{y}_2)$ or greater than or equal to it. If the former, the final *X*-to-*Y* intensification can be made up of the *Y*-to-*X* intensification whose magnitude is the left-hand-side together with $\underline{y}_2$. (It is easy to show that the singleton of this *X*-to-*Y* intensification satisfies the three conditions of NCP(S).) If the latter, transfer $\mu(\underline{y}_2)$ to the left-hand-side. The new left-hand-side now also represents a *Y*-to-*X* intensification. Turn to $\mu(\underline{y}_3)$. Either the new left-hand-side is less than $\mu(\underline{y}_3)$ or greater than or equal to it. If the former, the final *X*-to-*Y* intensification can be made up of the *Y*-to-*X* intensification represented by the new left-hand-side together with $\underline{y}_3$. If the latter, transfer $\mu(\underline{y}_3)$ to the left-hand-side. This process can be continued until a final *X*-to-*Y* intensification is found.

r.22

$\underline{x}_m \neq \underline{x}_n$

From (*r*) and the fact that $\sum_{i=1}^{m} \mu(\underline{x}_i) \geq \mu(\underline{y}_1)$ (from the definition of '$\underline{x}_m$'), we can derive the fact that:

$$\sum_{i=m+1}^{n} \mu(\underline{x}_i) < \sum_{j=2}^{r} \mu(\underline{y}_j)$$

But this has the same form as the main assumption of case (*r* - 1) (i.e. $\sum_{i=1}^{n} \mu(\underline{x}_i) < \sum_{j=1}^{j=r-1} \mu(\underline{y}_j)$), since the number of $\mu(\underline{y}_j)$ s on the right-hand-side is *r* - 1. This means, on the hypothesis of this induction step that the thesis we are proving holds in case r - 1, that a final X-to-Y intensification can be constructed for it. Call this intensification 'I'. What we hope to do, then, is use I, in conjunction with the Y-to-X intensification formed from the $\mu(\underline{x}_i)$ s from 1 through m to form the final X-to-Y intensification for case (r).

From (r) we can get:

$$\text{(iii)}\ \mu(\underline{x}_m)-\left(\mu(\underline{y}_1)-\sum_{i=1}^{m-1}\mu(\underline{x}_i)\right)<\sum_{j=2}^{r}\mu(\underline{y}_j)-\sum_{i=m+1}^{n}\mu(\underline{x}_i)$$ [9]

We know from previous reasoning concerning $\underline{x}_m$ that the left-hand-side of (iii) must be the magnitude of a Y-to-X intensification. Now since I is the final X-to-Y intensification for an instance of case (r - 1) involving X-segments $\underline{x}_{m+1} \ldots \underline{x}_n$, it must contain all these X-segments as ultimate constituents. Furthermore, since it is an *X-to-Y* intensification, the magnitudes of these X-segments must be *subtracted* and not added in the formula giving its magnitude. For the same reason, the magnitudes of the Y-segments must be *added* and not subtracted. So far the properties of I agree with the requirements of the right-hand-side of (iii). The only snag is that there is no reason to suppose that I must involve *all* r - 1 Y-segments. Suppose I is such that it comes under case (r - 1)'s equivalent of case r.21 above (i.e. case (r - 1).21). There, as the r - 1 equivalent of (ii) is progressively modified, the final X-to-Y intensification is formed as soon as we hit on a Y-segment whose magnitude is greater than the left-hand-side at that point, all *subsequent* *Y*-segments being ignored. Thus we have to allow for the possibility that not all the *Y*-segments from $\underline{y}_2$ through $\underline{y}_r$ are represented in *I*. Let us suppose that they are indeed not all represented. (If they *are*, $\mu(\underline{I})$ completely coincides with the right-hand-side of (iii) and the construction of the final *X*-to-*Y* intensification for case (*r*) is straightforward.) Since $\mu(\underline{I})$ is given by the sum of *some* of the $\mu(\underline{y}_j)$ s from 2 through *r* minus all the $\mu(\underline{x}_i)$s from m + 1 through n, the right-hand-side of (iii) must be equal to:

$$\mu(\underline{I}) + \sum^{(-I)}\mu(\underline{y}_j)$$

where ' $\sum^{(-I)}\mu(\underline{y}_j)$' is intended to represent the sum of all the $\mu(\underline{y}_j)$s from $\underline{y}_2$ through $\underline{y}_r$ not occurring as ultimate constituents of *I*. Thus (iii) can be re-written as:

(iii') $\mu(\underline{x}_m) - \left(\mu(\underline{y}_1) - \sum_{i=1}^{m-1}\mu(\underline{x}_i)\right) < \mu(\underline{I}) + \sum^{(-I)}\mu(\underline{y}_j)$

Now it may be that the left-hand-side of (iii') is less than $\mu(\underline{I})$ *alone*. If it is, then the method of construction of the final *X*-to-*Y* intensification for case (*r*) is obvious. If it is *not*, then the left-hand-side minus $\mu(\underline{I})$ represents a *Y*-to-*X* intensification. Now (iii') entails

(iii'') $\mu(\underline{x}_m) - \left(\mu(\underline{y}_1) - \sum_{i=1}^{m-1}\mu(\underline{x}_i)\right) - \mu(\underline{I}) < \sum^{(-I)}\mu(\underline{y}_j)$

Consider the first of the $\mu(\underline{y}_j)$ s on the right-hand-side of (iii''). Either it is greater than the left-hand-side or less than or equal to it. If the former, then the construction of the final *X*-to-*Y* intensification for case (*r*) is now obvious. If the latter, transfer this $\mu(\underline{y}_j)$ to the left-hand-side. Now consider the *second* $\mu(y_j)$ on the right-hand-side. Either it is greater than the new left-hand-side or less than or equal to it. If the former, then the construction of the final *X*-to-*Y* intensification for case (*r*) is now obvious. If the latter, transfer this second $\mu(\underline{y}_j)$ to the left-hand-side. Clearly this process is bound to lead eventually to a final *X*-to-*Y* intensification for case (*r*).

This completes the proof of E1 and with it that of the entire Equivalence Thesis.

ENDNOTES

[1] NCP(S) was of course stated in its A-version, presupposing temporal discreteness. It would probably be possible to state an appropriate Version B, and presumably the proof of the equivalence of this principle to the Total Disutility Principle would not be radically different from the proof that appears here.

[2] For the statement of NCP(S) and the definition of an *X*-to-*Y* intensification, see, respectively, pp. 110 and 108-109.

[3] To be more explicit: a non-zero *X*-to-*Y* intensification is either of order zero or of order greater than zero. In the former case, it is simply a *Y*-segment; in the latter, it corresponds to one of the statements described in the text, and the '$\leq$' in that statement can be replaced by a '$<$'.

[4] Given the assumption of temporal discreteness that is operating here (see note 1), this can be done for any instance of case (a).

[5] Strictly speaking, it is not entirely clear that

such a segmentation can always be produced. One problem is that the minimal units of experience might not always be of the same duration. But this problem could be overcome by invoking <u>hypothetical</u> divisions of minimal units. However, there is a further difficulty. Suppose the duration of a minimal unit is an irrational number. In that case, even an adequate hypothetical division is not possible. But this is not really a major problem. We can always come up with a hypothetical division with segments that are equal in duration to any desired level of accuracy, and this is all that is really necessary. (I am grateful to David Hitchcock for drawing these points to my attention.)

[6] Note that the main assumption of each case is given the same number as that case itself.

[7] Strictly, this assumes that $n > 1$, which may not be the case. If it is not, the statement $\mu(\underline{x}_n) \geq \mu(\underline{y}_1)$ may be used without fundamentally affecting what follows.

[8] Strictly speaking, there would be no $\sum_{i=1}^{n-1} \mu(\underline{x}_i)$ in the case where n = 1. For that case, this expression can simply be omitted without fundamentally affecting the arguments that follow.

[9] The same comment as in the previous note applies to $\sum_{i=1}^{m-1} \mu(\underline{x}_i)$ with respect to the case where m = 1.

APPENDIX B

THE NON-TRANSITIVITY OF CORRELATIVE SUPERIORITY

Suppose situations A_1, A_2 and A_3 each have minutes of suffering with intensities as given below:

A_1	A_2	A_3
20	11	6
		6
	10	6
		5

By NCP2, A_2 is worse than A_1, since the 20 can be mapped onto the 11 and the resulting pair traded off against the 10 (since 10 is greater than 20 - 11). But in precisely the same way, the 11 and 10 in A_2 can be countered, respectively, with the first two 6s and with the third 6 and the 5 in A_3. Thus A_3 is also worse than A_2 by NCP2. Clearly by repeating this operation as many times as needed, we can get down to a situation A_n in which all the segments are of trivial intensity compared with the 20 in A_1. But by INCPO, A_n will be a much better situation than A_1. In other words, we will have a chain of possible situations A_1, A_2 . . . A_n, such that each member is worse than its predecessor, but such that the first member A_1 is much worse than the last member A_n. This is somewhat surprising.

We can make the point more obvious by increasing the *degree* of NCP2-superiority between successive members of the chain. In the above case, the difference was only a matter of 1 unit in each of the final trade-offs (i.e. in the one final trade-off from A_1 to A_2 and in each of the two final trade-offs from A_2 to A_3). In the following case, it is 9:

B_1	B_2	B_3
20	19	18
		10
	10	9
		10

Again, by continuing this process as long as we need to, we could eventually get to a situation B_n in which all the segments were of merely trivial intensity and which was thus clearly superior to B_1. The process would be a longer one than in the previous case, but it could still be achieved. And because the *degree* of NCP2-inferiority between each member of the chain and its predecessor is so great, it is even harder to doubt now that it really *is* inferior.

The phenomenon we have encountered here results from the fact that the relation of correlative superiority is non-transitive. Transitivity may not fail in the sense that there are cases in which X is definitely superior to Y and Y is definitely superior to Z and yet X is definitely *not* superior to Z. But it does fail in the sense of there being cases in which X is less clearly and determinately superior to _Z_ than _X_ is to _Y_ and _Y_ is to _Z_. Thus in the example on p. 159, $\underline{A}_1$ is superior to $\underline{A}_2$ by NCP2 and $\underline{A}_2$ is superior to $\underline{A}_3$ by NCP2. But $\underline{A}_1$ is only superior to $\underline{A}_3$ by NCP4, which is a much less acceptable principle than NCP2. Thus $\underline{A}_1$ is much less clearly and determinately superior to $\underline{A}_3$ than it is to $\underline{A}_2$ and $\underline{A}_2$ is to $\underline{A}_3$.

However, there is really no reason why we should expect the relation of correlative superiority to be transitive. For it is of a wholly different character from those conventional kinds of superiority-relations which _do_ guarantee transitivity. These conventional relations are based on the attribution to situations of 'absolute values' (or 'disvalues'). One situation is superior to another when it has a higher absolute value (or lower absolute disvalue). But if _X_ has a higher absolute value than _Y_ and _Y_ has a higher absolute value than _Z_, then X must have a higher absolute value than _Z_ and so transitivity is guaranteed. Conventional theories proceed as if there were literally one thing that each situation had a determinate amount of, and such that the superiority of _X_ over _Y_ depends on a comparison of the amounts that each have. But the present theory involves a different approach. '_Y_ is negatively utilitarianly worse than _X_' is not logically founded on a judgement of the form '_X_ has amount of suffering _n_ and _Y_ has amount of suffering _m_, such that _m_ is greater than _n_', but on considerations of an entirely different sort having to do with certain kinds of relationships between the components of _X_ and _Y_.

Failure of transitivity cannot be used as a convincing _objection_ to the present theory. For our intuitions clearly show that the relation of utilitarian superiority _is_ non-transitive, and the theory merely reflects those intuitions. Thus in the second of the two examples, it is very hard to resist the suggestion that $\underline{B}_1$ is superior to $\underline{B}_2$, $\underline{B}_2$ to $\underline{B}_3$ and so on, until we get to $\underline{B}_n$. But it is also extremely hard to resist the suggestion that $\underline{B}_n$ is vastly superior to $\underline{B}_1$. The intuitions are very firm and they concern relatively uncomplex matters. Thus it is hard to believe that they could involve a logical error.

BIBLIOGRAPHY

Adams, Marilyn McCord. "Hell and the God of Justice", Religious Studies, 11 (1975), 433-447.

Bergström, Lars. "Interpersonal Utility Comparisons", Grazer Philosophische Studien, 16/17 (1982), 283-312.

Brandt, R. B. "Problems of Contemporary Utilitarianism: Real and Alleged". Ethical Theory. Edited by Norman E. Bowie. Indianapolis/Cambridge: Hackett Publishing Company, 1983.

-------------. A Theory of the Good and the Right. Oxford: Clarendon Press; New York: Oxford University Press, 1979.

Broad, C. D. Five Types of Ethical Theory. London: Routledge & Kegan Paul Ltd., 1930.

Brook, Richard and Schwimmer, Seymour. "On Adding the Good", Social Theory and Practice, 7 (Fall, 1981), 325-335.

Hare, R. M. Freedom and Reason. Oxford: Clarendon Press, 1963.

-----------. Moral Thinking: its Levels, Method and Point. Oxford: Clarendon Press, 1981.

Harsanyi, John. "Morality and the Theory of Rational Behaviour". Utilitarianism and Beyond. Edited by Amartya Sen and Bernard Williams. Cambridge: Cambridge University Press, 1982. Paris: Editions de la Maison des Sciences de l'Homme, 1982.

Hurka, T. M. "Average Utilitarianisms", Analysis, 42 (March, 1982), 65-69.

-----------. "More Average Utilitarianisms", Analysis, 42 (June, 1982), 115-119.

Kavka, Gregory. "The Numbers Should Count", Philosophical Studies, 36 (October, 1979), 285-294.

McDermott, M. "Utility and Distribution", Mind, 91 (October, 1982), 572-578.

Ng, Yew-Kwang and Singer, Peter. "An Argument for Utilitarianism", Canadian Journal of Philosophy, 11 (June, 1981), 229-239.

Parfit, Derek. "Innumerate Ethics", Philosophy and Public Affairs, 7 (Summer, 1978), 285-301.

-------------. Reasons and Persons. Oxford: Clarendon Press, 1984.

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

Regan, Donald. Utilitarianism and Co-operation. Oxford: Clarendon Press; New York: Oxford University Press, 1980.

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

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

Taurek, John. "Should the Numbers Count?", Philosophy and Public Affairs, 6 (Summer, 1977), 293-316.